Capítulo 1

Introdução à Mineração de Opiniões

Karin Becker

Programa de Pós-Graduação em Ciência da Computação - Instituto de Informática
Universidade Federal do Rio Grande do Sul (UFRGS)
karin.becker@inf.ufrgs.br

Abstract

Opinions are central to almost all human activities and are key influencers of our behavior. Governments, companies and organizations rely on public opinion to define strategies to improve the services they provide, or increase the success and visibility of the brands, entities or causes they represent. Social networks, forums, twitter, on-line newspapers and sites for the evaluation of products and services are examples of the many platforms available in the web, which allow users to express their opinions. Opinion mining, or sentiment analysis, is a recent field that aims at automatically identifying opinionative content, and determine the sentiment, perception or attitude of the public with regard to the target of the opinion. This chapter aims to introduce the main concepts and techniques used for opinion mining, and illustrate their application in a case study and related work. It comprises the motivation for the field; the theoretical foundation and fundamental concepts; the basic natural language processing functions required; the opinion mining process and the polarity classification approaches; an illustration using a real case study; and a discussion of related work.

Resumo

Opiniões são fundamentais para quase todas as atividades humanas, e influenciam sobremaneira nosso comportamento. Governos, empresas e organizações dependem de opinião pública para definir estratégias que melhorem os serviços que prestam, ou aumentem o sucesso e visibilidade das marcas, entidades ou causas que representam. Redes sociais, fóruns, twitter, jornais e sites de avaliação de produtos e serviços on-line, são

exemplos das plataformas web que permitem a usuários expressarem suas opiniões. A mineração de opinião, ou análise de sentimento, é um campo recente que visa identificar automaticamente o conteúdo de opinião, e determinar o sentimento, percepção ou atitude do público em relação a seu alvo. Este capítulo tem como objetivo apresentar os principais conceitos e técnicas utilizadas para a opinião de mineração, e ilustrar sua aplicação prática usando um estudo de caso, bem como trabalhos relacionados. Ele compreende a motivação para a área, a fundamentação teórica e os conceitos fundamentais; as funções básicas de processamento de linguagem natural necessárias; o processo de mineração de opinião e as abordagens de classificação de polaridade; uma ilustração usando um estudo de caso real, e uma discussão de trabalhos relacionados.

1.1. Introdução

Opiniões são fundamentais, e têm grande influência sobre o comportamento das pessoas. Saber a opinião de outros constitui um fator crítico na tomada de decisões. Conversas com amigos ou pessoas próximas podem influenciar decisões simples como qual filme iremos assistir ou onde tiraremos férias. A opinião de especialistas ou a leitura de publicações especializadas podem embasar decisões mais complexas, como qual carro comprar, quais os melhores investimentos ou a escolha de nossos representantes políticos. Organizações também baseiam suas estratégias de negócios, investimentos ou marketing na opinião de seus clientes sobre seus produtos ou serviços. A mensuração da satisfação dos clientes é constante em ambientes de comércio (tradicionais ou virtuais). Empresas que desejam lançar um novo produto, ou definir novas estratégias de atuação no mercado, estão interessadas em avaliar a aceitação junto a seu público alvo. Políticos definem suas estratégias de campanha baseados na opinião dos eleitores.

A importância da opinião é tão grande que muitas empresas têm seu negócio voltado à obtenção deste tipo de informação. Tradicionalmente, a resposta a questões envolvendo a opinião pública envolve técnicas como pesquisa de campo, telefonemas ou preenchimento de questionários. Estas técnicas envolvem custos, são restritas a um grupo de foco bem definido ou amostra, e seu retorno é demorado e, por vezes, pouco eficaz. A latência da opinião também é alta, devido ao longo tempo necessário entre a coleta dos dados brutos, e disponibilização dos resultados de sua análise.

A explosão das mídias sociais alterou este cenário, disponibilizando a indivíduos e organizações conteúdo de opinião diversificado e em grandes volumes. Usuários da web têm a oportunidade de divulgar suas ideias e opiniões através de comentários, fóruns de discussão, blogs, *tweets*, redes sociais, entre outros. Isto aumenta as opções dos indivíduos na busca de opiniões, pois não estão mais limitados a sua rede pessoal de contatos (e.g. familiares, amigos, conexões profissionais) ou a opiniões de especialistas disponíveis publicamente (e.g. revistas, jornais). No tocante às organizações, isto significa oportunidades de ampliar as fontes de opinião quantitativa e qualitativamente, tornar mais baratas as formas de coleta, bem como reduzir o tempo necessário para disponibilização da informação. Contudo, o grande volume de informação produzido diariamente implica a necessidade de métodos e ferramentas capazes de processar automaticamente não apenas o conteúdo das publicações, mas também a opinião e sentimento que expressam.

A *mineração de opiniões*, também chamada de *análise de sentimento* ou *análise de subjetividade* [Liu 2012, Tsytsarau and Palpanas 2012, Pang and Lee 2008], é uma disciplina recente que congrega pesquisas de mineração de dados, processamento de linguagem natural, recuperação de informações, inteligência artificial, entre outras. A mineração de opinião é definida em [Liu 2010] como qualquer estudo feito computacionalmente envolvendo opiniões, sentimentos, avaliações, atitudes, afeições, visões, emoções e subjetividade, expressos de forma textual. O problema da mineração de opiniões pode ser estruturado em termos das seguintes tarefas genéricas [Tsytsarau and Palpanas 2012]: a) identificar conteúdo subjetivo sobre determinado assunto ou alvo em um conjunto de documentos; b) classificar a polaridade desta opinião, isto é, se tende a positiva ou negativa; e c) apresentar os resultados de forma agregada e sumarizada. A polaridade da opinião define o sentimento, percepção ou atitude do público em relação ao alvo da opinião.

Uma das aplicações pioneiras da mineração de opiniões é a extração e sumarização automática de opiniões a partir de revisões *on-line* de produtos e serviços [Pang et al. 2002, Hu and Liu 2004, Turney 2002, Dave et al. 2003, Ghani et al. 2006, Liu et al. 2013, Tang et al. 2009]. Além da integração automática de opiniões de várias fontes, e da sumarização de opiniões, outros problemas que podem ser citados nesta categoria de aplicações são a detecção automática de revisões falsas (i.e. com objetivo de autopromoção, ou de prejudicar o concorrente), e a correção de revisões mal classificadas [Liu 2012]. O uso de opiniões para alavancar funcionalidades de outros tipos de sistemas também tem um enorme potencial. Por exemplo, sistemas de recomendação poderiam recomendar produtos com avaliações positivas [Chen and Wang 2014, Sohail et al. 2013]; a colocação dinâmica de anúncios poderia detectar cenários favoráveis ao produto do anúncio [Fan and Chang 2010]; a análise de citações poderia detectar se trabalhos estão sendo citados como referências do estado da arte ou como limitações, entre outros.

Outro tipo de aplicação bastante comum é a monitoração de entidades específicas (e.g. políticos, celebridades, marcas) em redes sociais [Pak and Paroubek 2010, Calais Guerra et al. 2011, Castellanos et al. 2011] ou notícias [Godbole et al. 2007]. Estas são utilizadas para marketing, construção de marcas, relacionamento com clientes, entre outros propósitos. Várias ferramentas foram desenvolvidas com este objetivo, entre elas TweetSentiments[1], Sentimonitor[2], UberVU[3], BeOnPop[4], etc. A mineração de opinião sobre textos menos estruturados, como notícias e blogs, também tem sido alvo de bastante atenção [Balahur et al. 2009b, Balahur et al. 2010, Ku et al. 2006], as quais podem ser a base de sistemas de sistemas analíticos e de inteligência em *e-goverment*, *e-health*, etc. Outra vertente importante é o desenvolvimento de modelos preditivos baseados em sentimentos. Exemplos são a previsão de resultados de eleições [O'Connor et al. 2010, Tumasjan et al. 2010] ou de pesquisas de intenção de voto [Tumitan and Becker 2014], variações no mercado de ações [Bollen et al. 2011, Liu et al. 2013], preços e bilheteria de filmes [Asur and Huberman 2010, Archak et al. 2007], etc.

O propósito deste capítulo é apresentar os conceitos subjacentes à mineração de opiniões, e caracterizar cada uma das etapas do processo, descrevendo os problemas en-

[1] twittersentiment.appspot.com
[2] www.sentimonitor.com
[3] www.ubervu.com
[4] beonpop.com/pt

volvidos, e as técnicas que podem ser utilizadas. Etapas, problemas e técnicas serão ilustrados através de um estudo de caso. Também discutiremos trabalhos representativos encontrados na literatura, considerando diferentes tipos de texto: revisões de produtos, notícias e mídias sociais.

O restante deste capítulo está estruturado como segue. Na Seção 1.2 é descrita a fundamentação teórica, com o detalhamento dos conceitos que caracterizam a opinião, dos diferentes níveis de análise de opinião, e as diferentes formas de expressão de opinião. A Seção 1.3 discute a relação entre a mineração de opiniões e o processamento de linguagem natural (PLN), apresentando de forma pragmática como seus recursos fundamentam o processamento de texto com opiniões. A Seção 1.4 descreve o processo de mineração de opiniões em termos de suas principais etapas: identificação, classificação de polaridade e sumarização. As diferentes abordagens para a classificação de polaridade são discutidas na Seção 1.5. A Seção 1.6 ilustra em estudo de caso real, como as diferentes técnicas apresentadas foram empregadas, as dificuldades enfrentadas para a condução da mineração de opiniões, e as decisões tomadas. O uso da mineração de opiniões sobre diferentes tipos de dados é então discutida através de exemplos na Seção 1.7, considerando revisões de produtos, notícias e mídias sociais. Finalmente, a Seção 1.8 apresenta conclusões e perspectivas de pesquisa.

1.2. Fundamentação Teórica

1.2.1. Definições

A mineração de opiniões opera sobre porções de texto de quaisquer tamanho e formato, tais como páginas web, posts, comentários, *tweets*, revisões de produto, etc. Toda opinião é composta de pelo menos dois elementos chave: um *alvo* e um *sentimento* sobre este alvo [Liu 2012]. Um alvo pode ser uma entidade, um aspecto de uma entidade, ou um tópico, representando um produto, uma pessoa, uma organização, uma marca, um evento, etc. Já um sentimento representa uma atitude, opinião ou emoção que o autor da opinião tem a respeito do alvo. A *polaridade* de um sentimento corresponde a um ponto em alguma escala que representa a avaliação positiva, neutra ou negativa do significado deste sentimento [Tsytsarau and Palpanas 2012].

Mais formalmente, uma opinião corresponde a uma quíntupla $(\mathbf{e}_i, \mathbf{a}_{ij}, \mathbf{s}_{ijkl}, \mathbf{h}_k, \mathbf{t}_l)$ [Liu 2010], onde:

$\mathbf{e}_i$: é o nome de uma entidade;

$\mathbf{a}_{ij}$: é um aspecto da entidade $\mathbf{e}_i$ (opcional);

$\mathbf{s}_{ijkl}$: é a polaridade do sentimento em relação ao aspecto $\mathbf{a}_{ij}$ da entidade $\mathbf{e}_i$;

$\mathbf{h}_k$: é o detentor do sentimento (i.e. quem expressou o sentimento), também chamado de fonte de opinião;

$\mathbf{t}_l$: é o instante no qual a opinião foi expressa por $\mathbf{h}_k$.

O conceito de *aspecto*, também denominado característica (*feature*) ou propriedade, permite que uma entidade seja vista através de diferentes perspectivas ou atributos,

ou como uma hierarquia de partes e subpartes [Liu 2010]. Por exemplo, considere o comentário abaixo sobre um hotel, retirado do *site* Booking.com. Os aspectos deste hotel são o quarto, a vista, e o *wi-fi*. Assim, o sentimento expresso neste comentário é representado por quatro quíntuplas, sendo que uma refere-se ao hotel como um todo, e as demais, a aspectos específicos deste.

Cláudio - 31 de dezembro de 2013: "Adorei o hotel ***Vida Mansa****. Os* ***quartos*** *do hotel são super espaçosos, com uma* ***vista*** *linda para o mar. Pena que não há* ***wi-fi*** *nos quartos".*
– (Vida Mansa, geral, positivo, Cláudio, 31/12/2013)
– (Vida Mansa, quarto, positivo, Cláudio, 31/12/2013)
– (Vida Mansa, vista, positivo, Cláudio, 31/12/2013)
– (Vida Mansa, wi-fi, negativo, Cláudio, 31/12/2013)

Os termos *sentimento* e *opinião* frequentemente são usados como sinônimos neste contexto. A polaridade de um sentimento pode ser classificada em classes discretas (e.g. positiva, negativa ou neutra), ou como um intervalo que representa a intensidade deste sentimento, tipicamente [-1, 1]. Outra forma de expressão de sentimento é a *emoção*, usada para designar as percepções e pensamentos subjetivos de uma pessoa, tais como raiva, desgosto, medo, alegria, tristeza e surpresa. Ao contrário de uma opinião, a emoção não representa necessariamente um posicionamento ou uma atitude em relação ao alvo, e portanto, estes termos não são considerados sinônimos.

1.2.2. Níveis de Análise Textual

A detecção do sentimento em um texto pode ocorrer em diferentes granularidades, sendo que a decisão do nível está sujeita ao contexto e aplicação. A análise pode ser em nível de [Liu 2012]:

- *Documento*: nesse nível, a tarefa é classificar se um documento, tratado como um todo, expressa um sentimento positivo ou negativo. Esta granularidade é adequada quando o documento trata de uma única entidade, por exemplo, um documento que forneça uma opinião sobre um dado produto;

- *Sentença*: determina o sentimento de uma sentença específica de um documento. Este nível é bastante utilizado quando um mesmo documento contém opiniões sobre várias entidades. Ele também permite identificar e distinguir sentenças objetivas (fatos) e subjetivas (opiniões). Alguns autores sugerem ir além do nível de sentença, dividindo-a em cláusulas (e.g. "A cidade é péssima, mas a população é muito simpática") [Thet et al. 2010];

- *Entidade e Aspecto*: este nível foca na opinião expressa, independentemente dos construtos utilizados para expressá-la (e.g. documentos, sentenças, orações). Neste caso, o alvo da opinião pode ser uma entidade, ou algum de seus aspectos. No exemplo "Adoro minha câmera X porque a qualidade de sua lente é excepcional. Pena que o preço seja tão alto", observa-se que existem três opiniões em 2 sentenças: sobre a câmera X, e sobre dois de seus aspectos (preço e lente). Apenas

a opinião sobre o preço é negativa, sendo que a opinião sobre a lente, e sobre a câmera em geral são positivas. Este é o nível mais complexo de análise, o qual tem sido bastante estudado no contexto de revisões de produtos e serviços (e.g. [Hu and Liu 2004, Thet et al. 2010, Ghani et al. 2006, Qiu et al. 2011, Liu et al. 2013]).

1.2.3. Expressão de Opiniões

Opiniões referem-se a conteúdo subjetivo, escrito em linguagem natural. A forma como as opiniões estão expressas influencia diretamente a habilidade de processá-las corretamente.

Opiniões podem ser *regulares* ou *comparativas*; *diretas* ou *indiretas*, e *implícitas* ou *explícitas*. Em opiniões regulares, o autor da opinião expressa um sentimento, atitude, emoção ou percepção sobre um alvo ("Este filme é muito bom"). Já as opiniões comparativas expressam o sentimento com base na relação de similaridades ou diferenças entre duas ou mais entidades, ou preferência quanto a algum aspecto compartilhado. Assim, em "O teclado deste telefone é melhor do que o do meu telefone antigo", a opinião só pode ser efetivamente conhecida após estabelecer a opinião sobre o referencial (i.e. teclado do telefone antigo). As opiniões podem ser diretas ou indiretas. Um exemplo de opinião direta é "Este remédio é muito bom". Já em "Minha gripe piorou depois que tomei este remédio", a opinião negativa sobre o remédio é implicada pelo seu efeito sobre a gripe. Finalmente, opiniões explícitas expressam diretamente o sentimento, enquanto que as implícitas sugerem-no indiretamente ("Formou-se um vale no colchão que comprei na semana passada"). A maioria dos trabalhos assume que opiniões são regulares, diretas e explícitas, por serem mais fáceis de serem tratadas. Os demais tipos de opinião necessitam de uma avaliação complexa sobre a semântica e pragmática do texto para determinação de seu sentido, como será discutido na Seção 1.3.

É comum o uso de palavras de sentimento (e.g. ótimo, detesto) para expressar opiniões, mas seu uso não é condição necessária, nem suficiente. Nem toda opinião é expressa com palavras de sentimento (e.g. "Comprei este casaco na semana passada, e já está cheio de bolinhas"). Da mesma forma, o uso de uma palavra de sentimento não implica opinião. Por exemplo, a frase "Ando procurando um bom livro" expressa o fato da procura. Mais ainda, subjetivamente, alguns poderiam inclusive interpretar que a opinião implicitamente representada é a dificuldade de encontrar um bom livro. É importante ressaltar também que as palavras de sentimento podem assumir sentidos específicos de acordo com o contexto. Por exemplo, a sentença "Este smartphone é muito caro" expressa uma opinião negativa usando o termo "caro", enquanto que em "Este amigo me é muito caro", o mesmo termo é utilizado no sentido positivo.

Finalmente, algumas opiniões são expressas com sarcasmo ou ironia, de forma implícita ou explícita, sendo necessária informação contextual para identificar que a opinião é exatamente oposta àquela sendo expressa. Por exemplo, a frase "Parabenizo os políticos brasileiros por toda a consideração que apresentam para com o povo e suas necessidades" pode estar significando exatamente o contrário do que está explicitamente dito.

1.3. Recursos Básicos de Processamento de Linguagem Natural

1.3.1. A Mineração de Opiniões e o Processamento de Linguagem Natural

A mineração de opiniões lida com o processamento de textos escritos em linguagem natural, com maior ou menor formalidade (e.g. jornais vs. *tweets*). Portanto, recursos de processamento de linguagem natural (PLN) são necessários como fundamento básico da mineração de opiniões. O PLN trata computacionalmente os diversos aspectos da comunicação humana, considerando formatos e referências, estruturas e significados, contextos e usos. Muitos dos desafios do PLN estão relacionados à compreensão da linguagem, oral ou escrita, a qual pode se dar em diferentes níveis [Jurafsky and Martin 2010]:

- *fonético e fonológico:* relacionamento das palavras com os sons que produzem;
- *morfológico:* construção das palavras a partir unidades de significado primitivas e como classificá-las em categorias morfológicas;
- *sintático:* relacionamento das palavras entre si, cada uma assumindo seu papel estrutural nas frases, e como as frases podem ser partes de outras, constituindo sentenças;
- *semântico:* relacionamento das palavras com seus significados e como eles são combinados para formar os significados das sentenças; e
- *pragmático:* uso de frases e sentenças em diferentes contextos, afetando o significado do discurso.

A Figura 1.1 apresenta a visão clássica dos principais estágios de processamento da linguagem natural. Estes estágios refletem as distinções da linguística teórica entre *Sintaxe*, *Semântica* e *Pragmática* [Indurkhya and Damerau 2010].

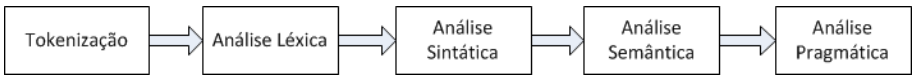


Figura 1.1. Estágios de análise no PLN.

Os três primeiros estágios têm por objetivo derivar a estrutura sintática das sentenças. O tokenizador tem por objetivo segmentar o texto em unidades menores denominadas *tokens*, que a grosso modo correspondem às palavras ou símbolos relevantes à linguagem. A segmentação envolve a descoberta de fronteiras entre palavras e entre sentenças, por vezes de forma entrelaçada, bem como ações de normalização de *tokens*. O analisador léxico tem por objetivo classificar *tokens* em diferentes categorias morfológicas relevantes. Finalmente, o analisador sintático (*parser*) trabalha em nível de agrupamento de palavras, analisando a constituição das frases de acordo com regras gramaticais. Dois tipos recursos são fundamentais nestes estágios: léxicos e gramáticas. De forma genérica, o propósito de um léxico é prover uma grande gama de informações sobre palavras, como etimologia, pronúncia, morfologia, sintaxe, entre outras. Já uma gramática define um conjunto de regras de boa formação das palavras e das sentenças de uma língua.

De certa forma, a granularidade dos três primeiros estágios reflete a maturidade do estado da arte em processar os aspectos sintáticos de uma linguagem. Muito menos se sabe sobre como associar significado às sentenças, tanto de forma independente do contexto (semântica), como em nível de discurso (pragmática) [Jurafsky and Martin 2010].

A análise de sentimentos é um problema de PLN, e como tal, sofre com a carência de soluções para várias questões, tais como tratamento de coreferências, negações, desambiguação, etc. Contudo, é um problema mais restrito, uma vez que não é necessária a compreensão completa do significado das sentenças e textos, mas sim das opiniões e seus alvos. Ainda, frequentemente a necessidade de desempenho para processamento de documentos em larga escala, e o uso de linguagens informais (e.g. próprias à internet) e/ou com erros previne a aplicação de técnicas mais sofisticadas.

Uma discussão mais ampla sobre PLN está fora do escopo deste capítulo, e pode ser encontrada em referências clássicas como [Indurkhya and Damerau 2010, Jurafsky and Martin 2010]. No restante desta seção discutiremos de forma pragmática alguns processamentos básicos subjacentes à mineração de opiniões, concluindo com uma breve discussão de alguns desafios que influenciam o processamento correto das opiniões.

1.3.2. Tokenização

A análise linguística de um texto exige o reconhecimento de caracteres, palavras e sentenças de um texto. O estágio de tokenização é fundamental para todas as tarefas de PNL, e de particular importância para a mineração de opiniões, porque a informação sobre sentimento é escassa, e por vezes expressa de modos particulares (e.g. :-) para representar alegria; S2 para representar afeição). Não há uma única maneira certa de reconhecer sentenças e *tokens*, e o algoritmo utilizado tem de levar em conta não apenas as características da linguagem, mas igualmente o objetivo da aplicação.

A tokenização pode ser precedida por algumas ações de pré-processamento no texto alvo. Por exemplo, em páginas web, pode ser interessante identificar e isolar certos rótulos HTML/XML, e substituir caracteres HTML por outras codificações padronizadas (e.g. substituir &lt pelo caracter Unicode <).

Os tokenizadores consideram regras específicas para fazer a segmentação de um texto em sentenças e em *tokens*. Por exemplo, o tokenizador baseado em *espaços em branco* coloca as palavras em letras minúsculas, e divide o texto ao reconhecer os seguintes separadores: espaço, tab, ou nova linha. Um tokenizador *Treebank style* utiliza as mesmas convenções do corpus Penn Treebank[5], um corpus público anotado sintática e semanticamente. As convenções do Treebank são as mesmas utilizadas em outros corpora de larga escala, e portanto, um padrão *de facto*. Neste tokenizador, a maioria dos caracteres de pontuação são utilizados como separadores. Em particular na língua inglesa, palavras como "can't" são transformadas nos *tokens* "can" e "'t", sendo dificultado O reconhecimento de uma negação (i.e. "can not" ou "cannot"). Outra consequência é que termos incluindo *hashtags* ou *emoticons* são desmembrados, assim como palavras escritas com hífen (além-mar) e nomes compostos (e.g. São Paulo).

Nem sempre as estratégias clássicas funcionam para recursos na internet e suas

[5]www.cis.upenn.edu/ treebank/home.html

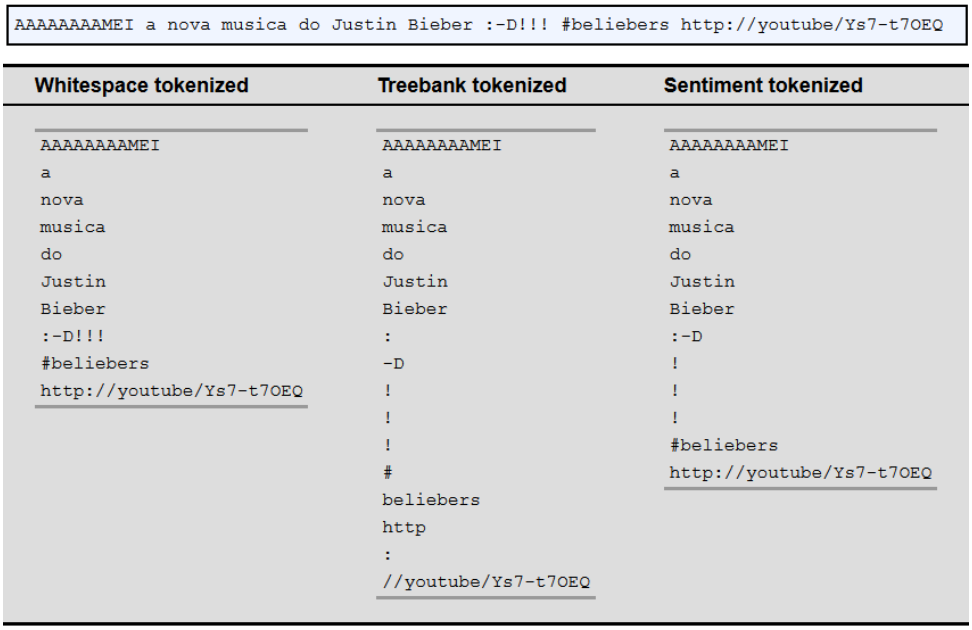


Figura 1.2. Comparação do resultado de três tokenizadores.

formas de expressão em geral, e de sentimento em particular. Cristopher Potts em [Potts 2011] propõe algumas heurísticas para tokenização no análise de sentimento, tais como: a) reconhecimento de *emoticons*; b) marcações específicas da internet (e.g. #hastags, @usernames, urls), c) xingamentos e palavrões (e.g. $#&!!, merd***), d) reconhecimento de siglas representando sentimento (e.g. LOL), entre outras. Os efeitos dos três tokenizadores discutidos podem ser comparados na Figura 1.2 para um *tweet* usando um script[6] disponibilizado por Potts em [Potts 2011] .

Outros aspectos importantes que podem ser lidados pelos tokenizadores usados no contexto da mineração de sentimentos são: a) tratar formas de intensidade, como por exemplo, o alongamento das palavras ("ameeeeeeei") e letras maiúsculas ("ODIEI esta música"); b) preservar nomes próprios ("Justin_Bieber"); c) identificar expressões idiomáticas. Potter resalta que o custo de algoritmos mais sofisticados de tokenização deve ser pesado face ao tamanho do corpus: quanto maior o corpus de treinamento para classificação da polaridade, menor o efeito da perda de *tokens* de sentimento não capturada por regras mais simples [Potts 2011].

Em termos de ferramentas, muitos tokenizadores estão disponíveis, quer como ferramentas ou funções isoladas (e.g. em Java, classe *StringTokenizer*, ou o método *String.split()*), ou como parte de um toolkit de processamento de linguagem natural (e.g. OpenNLP[7], Standford CoreNLP[8], NLTK[9]). Um toolkit particularmente interessante é o NLTK, na linguagem Python, que disponibiliza funcionalidades para muitas línguas, en-

[6] sentiment.christopherpotts.net/tokenizing
[7] opennlp.apache.org
[8] nlp.stanford.edu/software/corenlp.shtml
[9] www.nltk.org

tre elas, o português do Brasil. Outra opção para a língua portuguesa é o Palavras[10], uma solução bastante sofisticada, mas proprietária. LX_Center[11] e Linguateca[12] são sítios que agrupam recursos desenvolvidos para o tratamento da língua portuguesa.

1.3.3. Normalização: Radicais e Lemas

Alguns aspectos da normalização já foram mencionados como parte da tokenização, por exemplo, a transformação de letras maiúsculas em minúsculas, exceto quando se tratar de siglas (várias letras maiúsculas juntas, tal como "SBC") ou nomes próprios (palavras que iniciam com maiúsculo no meio da frase - “O congresso ocorreu em Brasília.”). Outras ações de normalização visam representar um grupo de palavras através de um único termo. Duas técnicas bastante usadas são o *stemming* e a lematização.

Stemming é um método para a redução de um termo ao seu radical, através da remoção de desinências, afixos, e vogais temáticas. Com sua utilização, todos termos derivados de um mesmo radical serão tratados como um único termo. Por exemplo, os termos "correr", "corrida", "corredor" possuem todos o mesmo radical: "corr". Note-se que um radical não é uma palavra que precisa existir. Cada algoritmo de stemming define um conjunto de regras para reduzir termos a seu radical, as quais são dependentes de língua. Por exemplo, uma das regras do *stemmer* Potter, voltado à língua inglesa, é eliminar todos os sufixos da palavras que terminam com "ing", desde que sejam precedidos por uma vogal. Assim, o radical de "moving" é "mov", mas o radical de "sing" é "sing". Outros exemplos de *stemmers* para a língua inglesa são o Lancaster e o WordNet. Para o português, exemplos de *stemmers* são o RSLP e o Snowball. *Stemmers* podem ser mais ou menos agressivos, de acordo com as regras que utilizam para eliminar os afixos.

Embora bastante usado na área de Recuperação de Informações para busca de documentos, este tipo de normalização não tem usualmente bons resultados na análise de sentimentos [Potts 2011]. Com efeito, por vezes a extração do radical elimina os afixos que diferenciam a polaridade de termos com o mesmo radical. Por exemplo, os termos "tolerância" e "tolerável" têm polaridades positiva e negativa, respectivamente, mas sua redução ao radical "toler" não permite esta diferenciação.

A lematização é o processo de agrupar as diferentes inflexões e variantes de um termo para que possam ser analisadas como um único item, o lema. O lema corresponde assim à representação canônica de um grupo de palavras, a qual é usada como entrada de verbete em dicionários. Por exemplo, "amor" é o lema de "amores", e "amar" é o lema de palavras como "amada" e "amarei". As diferenças podem ser resultado das conjugações dos verbos, flexões dos substantivos, ou declinações dos pronomes. Como técnica de normalização, a lematização é bem menos agressiva, mas possui um custo computacional alto por depender do conhecimento de classes gramaticais (vide Seção 1.3.4).

Os vários frameworks de PLN já citados na seção 1.3.2 possuem lematizadores e *stemmers*. Para a língua portuguesa, uma das melhores opções é, novamente, o NLTK.

[10] visl.sdu.dk
[11] lxcenter.di.fc.ul.pt
[12] www.linguateca.pt

1.3.4. Etiquetamento de Classes Gramaticais

Outro processamento importante no PLN é o etiquetamento de classes gramaticais, mais comumente designado por sua sigla em inglês *POS (Part of Speech)*. Este processo associa todas as palavras em um texto com suas respectivas classes gramaticais, baseado tanto na sua definição, quanto em seu contexto, isto é, relações com palavras adjacentes. As classes gramaticais cumprem muitos papéis no PLN, incluindo lematização e análise sintática. Especificamente na mineração de opiniões, palavras de sentimento são um forte indício da existência de sentimento em uma porção de texto, e portanto esta rotulação frequentemente é utilizada como informação de entrada à classificação da polaridade.

Para a rotulação, são utilizadas categorias de palavras ditas *abertas* e *fechadas*. As classes fechadas, tais como verbos auxiliares, pronomes, preposições, entre outros, incluem um número menor de palavras, que constituem um vocabulário mais estável. As classes abertas incluem os adjetivos, advérbios, substantivos, pronomes e interjeições, com um número maior de palavras, e vocabulário mais dinâmico. O etiquetamento POS necessita portanto um léxico com estas informações.

Técnicas para POS, e uma discussão mais completa sobre as questões de desambiguação e alinhamento, podem ser encontradas em [Jurafsky and Martin 2010]. Para nossos propósitos, é necessário compreender que um etiquetador POS toma como entrada uma série de rótulos, correspondendo às classes consideradas, e etiqueta os *tokens* de um texto de acordo com estes rótulos. A Figura 1.3 exemplifica o etiquetamento POS para a sentença "Amo usar camiseta branca" usando dois etiquetadores disponíveis no NLTK[13] e no LX_Center[14]. Em ambos os casos, é possível verificar as etiquetas V (Verbo), N (Substantivo) e ADJ (Adjetivo). No caso do LX_Center, o etiquetamento é mais rico. Por exemplo, junto com o verbo é detalhado o lema (e.g. Amar), o tempo verbal (e.g. presente do indicativo - pi), e conjugação (e.g. primeira pessoa do singular - 1s). Já para os substantivos e adjetivos, são informados, além do lema, o gênero e grau (e.g. fs - feminino e substantivo).

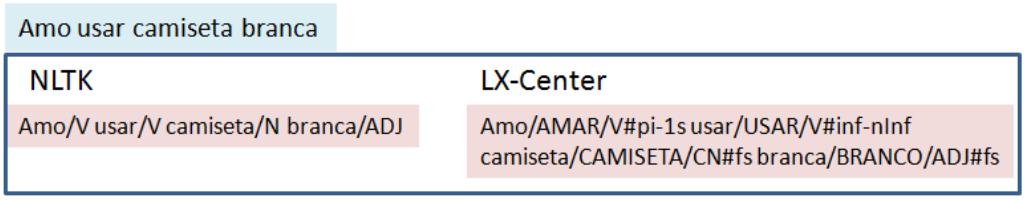


Figura 1.3. Comparação do resultado de duas marcações POS.

O etiquetamento POS é bastante explorado na análise de sentimento, porque é a forma mais básica de tratar a ambiguidade de palavras, com relativo baixo custo. Assume-se que a expressão de sentimentos está associada com certos padrões envolvendo as categorias das palavras. Por exemplo, [Turney 2002] busca locuções que contenham adjetivos e advérbios; [Hu and Liu 2004] utiliza substantivos para identificar os aspectos alvo de opiniões, e adjetivos para determinar a polaridade da opinião; [Mohammad et al. 2013]

[13] text-processing.com/demo/tag/
[14] lxcenter.di.fc.ul.pt/services/pt/LXServicesSuitePT.html

utiliza a frequência de cada classe de palavras para determinar a polaridade de *tweets*, entre outros exemplos.

1.3.5. Questões Avançadas

A análise sintática pode beneficiar a mineração de opiniões, derivando a estrutura de sentenças. Neste caso, um parser deriva uma árvore de dependências sintáticas. Alguns trabalhos utilizam estas técnicas para tratar opiniões em nível de cláusula, ou para identificar com maior precisão o alvo de uma opinião (e.g. [Qiu et al. 2011, Liu et al. 2013]). Uma árvore de dependências sintáticas é ilustrada na Figura 1.4. Ela contém informação que poderia ser usada para melhorar a precisão do da identificação do alvo da opinião (e.g. "good" é um modificador do substantivo "screen"), e tópicos ("screen" é um aspecto de "phone" pela ligação entre sujeito e predicado). Contudo, a área costuma adotar soluções mais pragmáticas devido à necessidade de processamento em larga escala [Castellanos et al. 2011, Godbole et al. 2007].

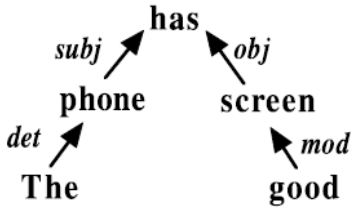


Figura 1.4. Exemplo de Árvore de Dependência Sintática (Fonte: [Liu et al. 2013]).

O tratamento da negação é uma questão importante na análise de sentimento, pois ela inverte o sentido das palavras usadas para designar uma opinião. A prática mostra que palavras com sentimento moderado como "bom" ou "ruim", ao serem negadas, têm sua polaridade invertida ("Esta câmera não é boa.", "Este filme não é ruim."). Contudo, a negação de palavras com forte conotação de subjetividade tem um espectro amplo de sentidos. Por exemplo, a sentença "Este filme não é excelente." não necessariamente significa que o filme é ruim [Potts 2011]. Outra questão é o escopo da negação, o qual não está limitado a palavras adjacentes. Métodos mais pragmáticos lidam com a questão rotulando com negação todos os *tokens* que seguem uma palavra de negação em uma oração [Pang et al. 2002, Potts 2011]. Por exemplo, em "Não sei se gostarei do filme: é meio parado.", o tokenizador pode modificar todos os termos da oração delimitada por :, ou seja, "sei_NEG","gostarei_NEG", "do_NEG", e "filme_NEG". Outra técnica possível é utilizar janelas de proximidade ao termo de negação (e.g. [Pang et al. 2002, Bollen et al. 2011]).

Ironia é uma forma negação bastante sutil, e constitui um dos problemas mais difíceis de se tratar na mineração de opiniões. O uso de sarcasmo é muito comum em alguns domínios, como discussões políticas e esportivas, opiniões sobre arte (filmes, bandas), etc [Sarmento et al. 2009, Turney 2002, Balahur et al. 2009a]. Alguns trabalhos encontrados na literatura para identificação de sarcasmo/ironia fazem uso de artifícios, como frequência do sinal de exclamação e interrogação, palavras capitalizadas, interjeições (e.g. "ah, oh, yeah"), *emoticons* e superlativos [Carvalho et al. 2009, Liu 2012].

Finalmente, outro desafio da PLN que impacta na mineração de opiniões é o de

co-referência, onde diferentes expressões em uma sentença ou documentos referem-se à mesma entidade. Por exemplo, as expressões “Dilma”, “Presidenta”, “Presidenta Dilma Rousseff” referem-se à mesma pessoa, devendo ser reconhecidas e unificadas. Uma forma de co-referência envolve o uso de pronomes, os quais devem ser resolvidos e relacionados com a respectiva entidade. Por exemplo, no texto “Paris é uma cidade maravilhosa. Ela é um excelente lugar para se visitar. Seus restaurantes são muito reconhecidos”, os pronomes “ela” e “seus” são co-referências de “Paris". O tratamento da co-referência é muito importante para a análise de sentimentos nos níveis de sentença e de aspectos, já que estes níveis analisam o sentimento de forma isolada (i.e. cada sentença ou opinião), com efeito direto sobre a revocação. Uma compilação de abordagens para resolução de co-referências é relatada em [Ng 2010]. Soluções para co-referência voltados à língua inglesa estão disponíveis nos toolkits OpenNLP e Standford CoreNLP. Heurísticas têm sido utilizadas em textos curtos como *tweets*, utilizando janelas de distância entre uma entidade nomeada e pronomes que as seguem [Castellanos et al. 2011].

1.4. Etapas da Mineração da Opinião

A mineração de opinião pode ser caracterizada em termos de três grandes tarefas [Tsytsarau and Palpanas 2012]: a) identificar (tópicos, sentenças opinativas), b) classificar a polaridade do sentimento, e c) sumarizar. Este processo é esboçado na Figura 1.5.

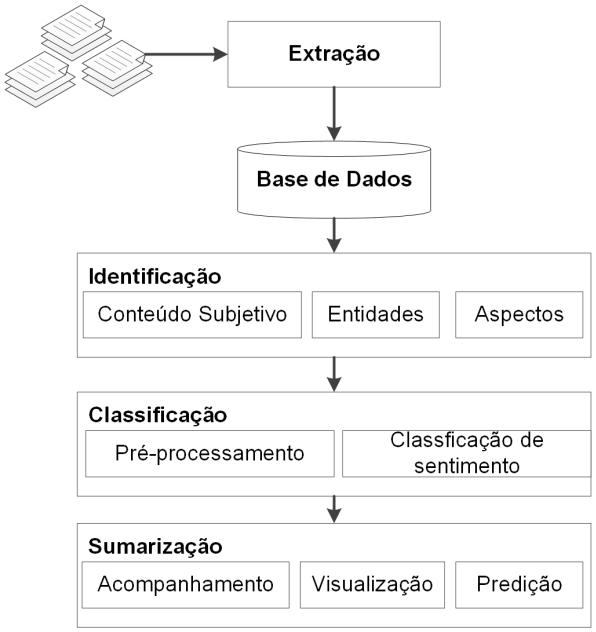


Figura 1.5. Etapas da Mineração de Opinião.

1.4.1. Identificação

Dado um conjunto de textos extraídos de alguma fonte (e.g. jornais, redes sociais, plataformas de revisão de produtos/serviços), a etapa de *identificação* consiste em encontrar os tópicos existentes (e possivelmente seus aspectos), e possivelmente associá-los com o respectivo conteúdo subjetivo. A forma de identificar as entidades, aspectos e sentimento são dependentes da granularidade escolhida para análise, e os algoritmos utilizados podem ser

distintos daqueles propostos para recuperação de documentos opinionativos [Pang and Lee 2008, Tsytsarau and Palpanas 2012].

A complexidade da identificação do alvo da opinião depende em muito da mídia considerada, e de seu grau de estruturação. A aplicação mais frequente em mineração de opiniões é a de revisão de produtos e serviços, porque o alvo pode ser mais facilmente identificado. Assume-se que todo o documento refere-se a uma única entidade, o alvo da revisão, sendo que o desafio está em identificar os aspectos desta entidade, se a análise for nesta granularidade.

Já em jornais, blogs ou posts, não se conhece *a priori* as entidades envolvidas, podendo inclusive envolver muitas entidades na mesma porção de texto. Na situação mais simples, pode-se restringir a identificação a entidades pré-definidas, como a busca de celebridades, atletas, políticos ou marcas. Um dos problemas neste caso é resolver os problemas de co-referência, já mencionados na Seção 1.3.5. Em mídias sociais, a co-referência pode ser um problema acentuado, pois as menções podem ser muito informais (apelidos, gírias com significado local ou temporal, *hashtags*, etc). Por exemplo, o termo "tricolor" no estado de São Paulo refere-se ao São Paulo Futebol Clube, enquanto que no estado do Rio Grande do Sul, esse termo designa o Grêmio Foot-Ball Porto Alegrense. Se a identificação não for direcionada a alvos pré-definidos, pode-se ainda utilizar técnicas de identificação de entidades nomeadas da recuperação de informações [Sarawagi 2008, Aggarwal and Zhai 2012].

Finalmente, esta tarefa pode envolver também o discernimento entre conteúdo ou sentenças com ou sem opinião, visando melhorar os resultados da próxima etapa. Isto é bastante comum quando o nível de análise é de granularidade menor. O critério utilizado para determinar o conteúdo de opinião é quase sempre a identificação de palavras de sentimento (e.g. "Eu recomendo este filme"), ou de classes de palavras candidatas a expressar sentimento (e.g. adjetivos).

1.4.2. Classificação da Polaridade

O problema de *classificação de sentimento*, também denominado *classificação de polaridade*, é frequentemente um problema de classificação binário, isto é, que classifica um dado texto em uma de duas classes: *positivo* ou *negativo*. No entanto, classes adicionais podem ser consideradas para que a análise seja mais robusta, ou para aumentar o nível de detalhe dos resultados. Assim, estas classes podem ser desdobradas em classificações com diferentes graus de intensidade (e.g. muitoPositivo, moderadamentePositivo), ou em intervalos numéricos representando um grau de intensidade [Tsytsarau and Palpanas 2012]. Neste último caso, a divisão do sentimento está relacionada à capacidade de definir algum limiar para distinguir os níveis de sentimento.

Outra abordagem é considerar a categoria neutra, que engloba textos sem uma tendência clara quanto a sua polaridade, ou simplesmente sem sentimento. Neste último caso, é a etapa de classificação de polaridade que tem como responsabilidade identificar textos sem sentimento de acordo com suas propriedades. Mas, como já mencionado, é mais frequente que este tipo de texto já tenha sido descartado na etapa anterior de identificação, porque a qualidade dos resultados da classificação costuma ser maior [Tsytsarau and Palpanas 2012].

Para a classificação da polaridade, diferentes abordagens são propostas na literatura, as quais são discutidas com maiores detalhes na Seção 1.5. Cada técnica pode necessitar de operações de pré-processamento e transformação específicas, tais como reconhecimento de construtos sintáticos, reconhecimento de n-gramas, extração de *features*, eliminação de termos irrelevantes, transformação em vetor de termos, etc.

Independente da abordagem empregada, a classificação da polaridade não é um problema trivial. Entre os principais desafios estão:

- o uso de palavras de sentimento pode ser enganoso e a polaridade de alguns termos são dependentes de contexto (conforme Seção 1.2.3) ;
- a dificuldade em tratar corretamente opiniões comparativas, implícitas e indiretas;
- muitos domínios são caracterizados pelo uso frequente de ironia e sarcasmo, onde o sentido implícito é exatamente oposto ao sentimento expresso explicitamente. Outros domínios (e.g. debates políticos, críticas culturais) estabelecem uma opinião positiva por oposição a uma argumentação negativa (ou vice-versa) [Balahur et al. 2009a, Pang et al. 2002, Turney 2002];
- a opinião pode depender do observador. Por exemplo, a opinião representada na sentença "Bom momento para as ações da Vale!" é positiva para quem detém este tipo de ação, mas pode ser péssima para quem deixou de investir nelas;
- a polaridade de conteúdo subjetivo nem sempre é objeto de consenso. Por exemplo, em anotações feitas por humanos, dificilmente o consenso é maior que 75% [Bruce and Wiebe 1999, Ku et al. 2006, Wiebe et al. 2005, Pang et al. 2002];
- a classificação é bastante dependente da extração das *features* do texto, a qual deve lidar com as várias questões da língua natural já discutidas na Seção 1.3.

1.4.3. Sumarização

Para poder identificar opinião média ou prevalecente de um grupo de pessoas sobre um determinado tópico/entidade, a opinião expressa por uma única pessoa não é suficiente, sendo necessário analisar uma grande quantidade de opiniões [Liu 2012]. É necessário a criação de métricas e sumários que quantifiquem a diversidade de opiniões encontradas a respeito um mesmo alvo. Este é o objetivo desta etapa, onde são criadas métricas que representem o sentimento geral, as quais podem ser visualizadas ou servir de entrada para outras aplicações.

Em revisão de produtos, um sumário de um determinado produto pode ajudar um consumidor a identificar seus respectivos pontos fortes e fracos, levando em consideração a experiência prévia de outras pessoas, expressas em suas opiniões. Esse tipo de recurso pode ser encontrado, por exemplo, no Google Shopping, que automaticamente extrai, analisa e agrega os aspectos de revisões de produtos disponibilizados por diferentes lojas de comércio eletrônico (e.g. Best Buy). A Figura 1.6 ilustra um exemplo de resultado desta ferramenta, onde aspectos como facilidade de uso, *design* e tamanho, foram identificados e exemplificados com uma sentença que demonstra o sentimento predominante sobre o

mesmo. Além disso, a ferramenta ainda mostra uma nota geral sobre o produto, baseada em uma classificação de estrelas.

Figura 1.6. Exemplo de sumarização de opiniões de aspectos produto extraídas de revisões (Fonte: Google Shopping).

Outra forma de sumarização, comum em aplicações que extraem de mídias sociais o sentimento do público em geral sobre uma determinada entidade (e.g. uma marca, produto, político, celebridade), é apresentar o sentimento na forma de relógios, ou associá-lo a informações temporais ou geográficas. Normalmente este tipo de mídia reflete o que as pessoas pensam sobre o alvo, dado algum evento. Por exemplo, o lançamento de um novo produto terá impacto nas redes sociais, que expressarão reações a esse acontecimento através de posts, comentários, *tweets*, endossos, etc. A empresa pode aproveitar-se disto para avaliar se este produto foi bem recepcionado pelo mercado. Um exemplo é a ferramenta UberVB, que sumariza e acompanha as menções e o sentimento em relação a uma determinada marca através do tempo. Os dados analisados são provenientes de várias mídias sociais, como o Twitter, Facebook, YouTube, blogs, etc. Na Figura 1.7 é mostrado o sentimento em relação à marca Microsoft, medindo-se o sentimento positivo, negativo e neutro através de um relógio, e bem como a tendência temporal do sentimento.

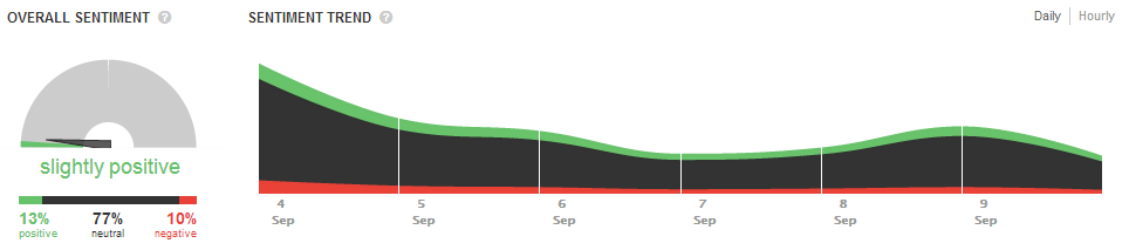


Figura 1.7. Sentimento extraído de mídias sociais em relação à Microsoft (Fonte: UberVU).

Outro exemplo de recursos de monitoração é a prova de conceito LCI [Castellanos et al. 2011], que monitora *tweets* sobre entidades definidas através de um conjunto de termos. LCI oferece um *dashboard*, ilustrado na Figura 1.8 para o filme “Piratas do Caribe". O *dashboard* inclui, da esquerda para a direita: a) uma árvore com os termos mais mencionados junto aos *tweets*; b) a freqüência destes termos usando uma nuvem de termos, e respectivo sentimento predominante, associando cores aos termos; c) um gráfico

de barras com o sentimento positivo, negativo e neutro dos termos mais frequentes; d) um gráfico de evolução de sentimento; e) um gráfico de evolução de volume de menções, e f) *tweets* representativos.

Figura 1.8. Dashboard para monitoração de sentimento do filme Piratas do Caribe (Fonte: [Castellanos et al. 2011]).

O sentimento sumarizado também pode ser utilizado para diversas aplicações, como prever eleições [Tumasjan et al. 2010], comportamento da bolsa de valores [Bollen et al. 2011, Liu et al. 2013], arrecadação de bilheterias de filmes [Asur and Huberman 2010], definição de preços [Archak et al. 2007], etc. No entanto, o sentimento puro (positivo ou negativo) pode não refletir de maneira correta o contexto analisado. Portanto, é importante criar métricas para representar o sentimento em relação ao alvo. Boa parte dos trabalhos na área utilizam a média do sentimento, ou a razão entre o sentimento positivo e negativo. Em certos casos, a predição pode ser feita somente com base na quantidade de menções às entidades, independente do sentimento sobre elas (e.g. [Tumasjan et al. 2010]).

Por exemplo, Social Market Analytics[15] é uma ferramenta para analisar a bolsa de valores, e ações de uma determinada companhia utilizando o Twitter. Na Figura 1.9 é mostrada a análise temporal do sentimento em relação às ações da *APPL - Apple Inc*. Esta ferramenta criou suas próprias métricas para análise, baseando na representação normalizada e ponderada da série de tempo do sentimento ao longo de um período retrospectivo (*S-Score*), e também uma média suavizada da métrica anterior (*S-Mean*). Estas métricas são comparadas com o valor de fechamento de uma determinada ação.

1.5. Abordagens de Classificação de Polaridade

As abordagens de classificação podem ser divididas em quatro grandes grupos: a) léxicas, com o uso de dicionários de sentimentos; b) aprendizado de máquina, com o uso predominante de técnicas de classificação; c) estatísticas, que valem-se destas técnicas para

[5] socialmarketanalytics.com

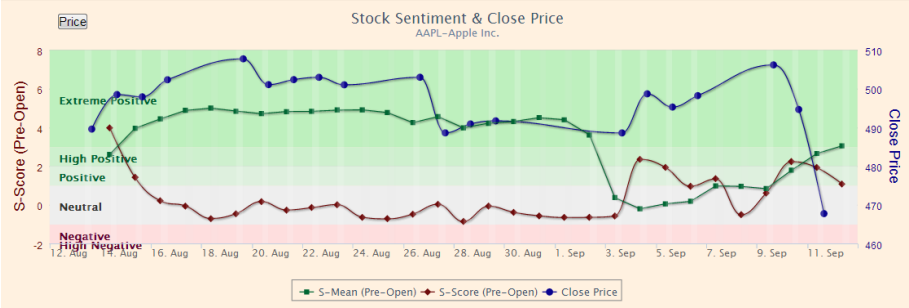


Figura 1.9. Relação entre sentimento e preço de ações (Fonte: Social Market Analytics).

avaliar a co-ocorrência de termos, e d) semânticas, que definem a polaridade de palavras em função de sua proximidade semântica com outras de polaridade conhecidas. Estas diferentes abordagens podem ser combinadas para melhoria de resultados (e.g. [Castellanos et al. 2011]). Uma revisão de trabalhos descrita em [Tsytsarau and Palpanas 2012] aponta uma predominâncias das duas primeiras abordagens.

1.5.1. Abordagem Baseada em Dicionário

A abordagem baseada em *dicionário* é também denominada *léxica* ou *linguística*. O aspecto central desta abordagem é o uso de léxicos (dicionários) de sentimentos, que são compilações de palavras ou expressões de sentimento associadas à respectiva polaridade.

1.5.1.1. Léxicos de Sentimentos

A composição básica de um léxico de sentimento é a palavra de sentimento e sua respectiva polaridade, expressa como uma categoria, ou como um valor em uma escala. Léxicos são relacionados a um idioma específico. Para a língua inglesa, podemos citar o General Inquirer [Stone et al. 1966], MPQA [Wiebe and Riloff 2005], SentiWordNet [Baccianella et al. 2010] e WordNetAffect [Strapparava and Valitutti 2004]. Já para a língua portuguesa estão disponíveis o OpLexicon [Souza et al. 2011] e o SentiLex-PT [Silva et al. 2012], sendo o primeiro para português do Brasil e o último, para português de Portugal. Alguns léxicos possuem versões para múltiplas línguas, como o Linguistic Inquiry and Word Counts (LIWC) [Tausczik and Pennebaker 2007], parte de um software de análise de texto desenvolvido para avaliar os componentes estruturais, cognitivos e emocionais de amostras de texto. Esforços estão sendo realizados para disponibilizar um versão do SentiWordNet para o português [de Paiva et al. 2012].

Existem muitas variação entre léxicos, incluindo o número de termos e as informações adicionais que são agregadas a suas entradas. Muitos dicionários possuem associadas a cada entrada informações adicionais, tais como: flexões (e.g. bonito, bonita, bonitos), o lema e/ou radical; POS, etc. A Tabela 1.1 resume as propriedades de alguns léxicos mencionados acima, em termos de número de entradas; disponibilidade de POS, radical e lema; e língua alvo.

As Figuras 1.10, 1.11 e 1.12 ilustram variações de três léxicos: MPQA, SentiWordNet e Sentilex-PT. No MPQA é possível verificar que os sentimentos são classi-

Tabela 1.1. Tabela comparativa de léxicos de sentimentos.

Dicionário	Pos	Neg	POS	Radical	Lema	Idioma
General Inquirer	1.915	2.291	S	N	N	Inglês
OpinionFinder	2.718	4.912	S	S	N	Inglês
OpLexicon	8.675	14.469	S	N	N	Português
SentiLex-PT	82.347 entradas		S	N	S	Português
SentiWordNet	117.659 entradas		S	N	S	Inglês

ficados como positivos e negativos, e modificados pelo nível de subjetividade forte ou fraca (*strength*). No SentiWordNet, a polaridade dos termos é representada por uma escala numérica, variando de 0 a 1 para cada tipo de polaridade (*PosScore* e *NegScore*). Note-se que o léxico inclui termos neutros (i.e. polaridade 0). O Sentilex-PT inclui tanto termos, quanto expressões idiomáticas, os quais estão relacionados a suas flexões e variações. A polaridade (*POL*) é positiva ou negativa (1 e -1, respectivamente), podendo haver variações se relacionada ao sujeito ou predicado da sentença.

	Strength	Length	Word	Part-of-speech	Stemmed	Polarity
1.	type=weaksubj	len=1	word1=abandoned	pos1=adj	stemmed1=n	priorpolarity=negative
2.	type=weaksubj	len=1	word1=abandonment	pos1=noun	stemmed1=n	priorpolarity=negative
3.	type=weaksubj	len=1	word1=abandon	pos1=verb	stemmed1=y	priorpolarity=negative
4.	type=strongsubj	len=1	word1=abase	pos1=verb	stemmed1=y	priorpolarity=negative
5.	type=strongsubj	len=1	word1=abasement	pos1=anypos	stemmed1=y	priorpolarity=negative
6.	type=strongsubj	len=1	word1=abash	pos1=verb	stemmed1=y	priorpolarity=negative
7.	type=weaksubj	len=1	word1=abate	pos1=verb	stemmed1=y	priorpolarity=negative
8.	type=weaksubj	len=1	word1=abdicate	pos1=verb	stemmed1=y	priorpolarity=negative
9.	type=strongsubj	len=1	word1=aberration	pos1=adj	stemmed1=n	priorpolarity=negative
10.	type=strongsubj	len=1	word1=aberration	pos1=noun	stemmed1=n	priorpolarity=negative

Figura 1.10. Porção do Léxico MPQA (Fonte: [Potts 2011]).

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	00001740	0.125	0	able#1	(usually followed by `to') having the necessary means or [...]
a	00002098	0	0.75	unable#1	(usually followed by `to') not having the necessary means or [...]
a	00002312	0	0	dorsal#2 abaxial#1	facing away from the axis of an organ or organism; [...]
a	00002527	0	0	ventral#2 adaxial#1	nearest to or facing toward the axis of an organ or organism; [...]

Figura 1.11. Porção do Léxico SentiWordNet (Fonte: [Potts 2011]).

1.5.1.2. Classificação da Polaridade

Um dos métodos mais utilizados na abordagem linguística é o da co-ocorrência entre alvo e sentimento, o qual não leva em consideração nem a ordem dos termos dentro de um documento (*bag-of-words*), nem suas relações léxico-sintáticas. Para a classificação do sentimento em um texto, basta que exista uma palavra de sentimento, onde sua polaridade é dada por um léxico de sentimentos. Esse método é extensamente empregado para o atrelamento de um sentimento a uma entidade em uma sentença. Por exemplo, na sentença "o iPhone é muito bom", a polaridade positiva da palavra "bom" é associada à entidade

```
aberração,aberração.PoS=N;FLEX=fs;TG=HUM:N0;POL:N0=-1;ANOT=MAN
bonita,bonito.PoS=Adj;FLEX=fs;TG=HUM:N0;POL:N0=1;ANOT=MAN
bonitas,bonito.PoS=Adj;FLEX=fp;TG=HUM:N0;POL:N0=1;ANOT=MAN
bonito,bonito.PoS=Adj;FLEX=ms;TG=HUM:N0;POL:N0=1;ANOT=MAN
bonitos,bonito.PoS=Adj;FLEX=mp;TG=HUM:N0;POL:N0=1;ANOT=MAN
engoliste em seco,engolir em seco.PoS=IDIOM;Flex=J2p|J2s;TG=HUM:N0;POL:N0=-1;ANOT=MAN
engolistes em seco,engolir em seco.PoS=IDIOM;Flex=J2p;TG=HUM:N0;POL:N0=-1;ANOT=MAN
engoliu em seco,engolir em seco.PoS=IDIOM;Flex=J4s|P3s;TG=HUM:N0;POL:N0=-1;ANOT=MAN
engulam em seco,engolir em seco.PoS=IDIOM;Flex=Y4p|S4p|S3p;TG=HUM:N0;POL:N0=-1;ANOT=MAN
engulamos em seco,engolir em seco.PoS=IDIOM;Flex=Y1p|S1p;TG=HUM:N0;POL:N0=-1;ANOT=MAN
```

Figura 1.12. Porção do Léxico SentiLex-PT

iPhone. O método por co-ocorrência apresenta bons resultados quando o nível de análise textual é de granularidade pequena, pois a palavra detentora do sentimento está próxima à entidade que qualifica. Sendo assim, este método é usualmente utilizado em análises de nível de sentença, cláusula ou até em documentos com poucos caracteres, como um *tweet*.

Quando aplicada em nível de maior granularidade, estabelece-se algum tipo de média sobre as palavras de sentimento encontradas. A Equação 1 mostra uma função genérica de determinação de polaridade em um documento D, onde S_w representa a polaridade de uma palavra w em um dicionário. A agregação pode levar em conta funções de peso e de modificação. A função $peso()$ pode ser, por exemplo, alguma medida de distância entre a palavra de sentimento e o alvo, ou de importância da palavra no texto (e.g. frequência). A função $modificador()$ pode ser usada para tratar negações, palavras de intensidade (e.g. muito), etc. Esta função de agregação também pode ser estendida a sentenças, cujas cláusulas podem combinar diferentes palavras de sentimento.

$$\mathrm{S(D)} = \frac{\sum_{w \in D} S_w \cdot peso(w) \cdot modificador(w)}{\sum peso(w)} \tag{1}$$

Existem métodos linguísticos mais complexos (e.g. [Liu et al. 2013, Qiu et al. 2011], que exploram a estrutura sintática do texto para aumentar a qualidade da classificação com base em informações morfossintáticas ali presentes (e.g. sujeito, predicado, dependências, etc.). No entanto, ferramentas de processamento de linguagem natural são em sua maioria restritas a determinado idioma, e portanto esta abordagem é mais utilizada em documentos na língua inglesa.

É importante ressaltar que, nesta abordagem, um texto neutro é diferente de um texto não polarizado. Um texto não polarizado é aquele no qual não há elementos suficientes para poder classificá-lo, e consequentemente, para o qual a tarefa de classificação não consegue chegar à conclusão sobre sua polaridade. Isso geralmente acontece quando o léxico é incompleto, ou quando o conteúdo analisado possui ruídos, tais como erros tipográficos e sentenças incompletas que impedem a verificação dos termos no dicionário [Dey and Haque 2009].

1.5.1.3. Discussão

Uma das grandes vantagens da classificação de polaridade baseada em léxico é o pré-processamento simplificado do documento, frequentemente reduzida à tokenização, normalização e POS. Palavras com baixo potencial de sentimento podem ser eliminadas, visando melhor desempenho de busca nos dicionários (e.g. *stop words*, categorias gramaticais como preposição e pronome, etc). Outra vantagem é que não é necessário haver um *corpus* rotulado como pré-condição da classificação da polaridade: basta procurar as palavras nos léxicos adotados.

Contudo, a maioria dos dicionários disponíveis são genéricos, e a polaridade do sentimento é bastante dependente de contexto. Ela também é sensível quando considerada sobre textos gerados por usuários em mídias informais (e.g. redes sociais, *tweets*), onde expressões regionais, gírias, abreviaturas típicas da internet, etc, são fartamente empregadas. Melhores resultados têm sido obtidos com dicionários dependentes de domínios [Hu and Liu 2004], criados de forma semi-supervisionada com base em métodos estatísticos de co-ocorrência (e.g. [Turney 2002], ou utilizando relações de sinônimos/antônimos de tesauros (e.g. [Hu and Liu 2004, Godbole et al. 2007]). Estas técnicas serão discutidas com mais detalhes na Seção 1.5.3. Contudo, são necessários métodos para verificação da sanidade de termos na criação de um novo vocabulário.

1.5.2. Abordagem baseada em Aprendizado de Máquina

O objetivo principal das técnicas de aprendizado de máquina é descobrir automaticamente regras gerais em grandes conjuntos de dados, que permitam extrair informações implicitamente representadas. Na área de mineração de opiniões, nota-se um predomínio do uso de métodos supervisionados de aprendizagem, mais especificamente, *classificação* e *regressão* [Tan et al. 2006].

1.5.2.1. Aprendizagem Supervisionada

O problema de classificação, assim como o de regressão, é dividido em dois passos, como ilustrado na Figura 1.13:

1. aprender o modelo: derivar um modelo de classificação sobre um conjunto de dados treino previamente rotulado com as classes consideradas (e.g. positivo, negativo);

2. aplicar o modelo: prever a classe de novas instâncias de dados utilizando o modelo preditivo resultante.

Os *dados de treino* para a classificação/regressão correspondem a um conjunto de registros caracterizadas por atributos. O rótulo é denominado atributo *classe* (ou *alvo*), enquanto que os demais são designados como atributos *discriminantes* ou *features*. O atributo alvo na classificação é discreto, enquanto que na regressão, ele é numérico.

No caso de classificação de textos, as *features* correspondem primariamente a representações de termos utilizados. A Figura 1.14 ilustra como um conjunto de 3 frases

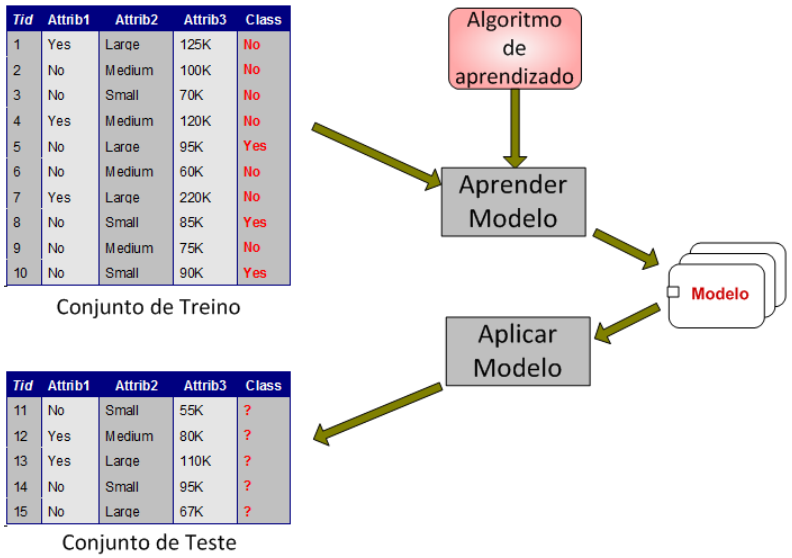
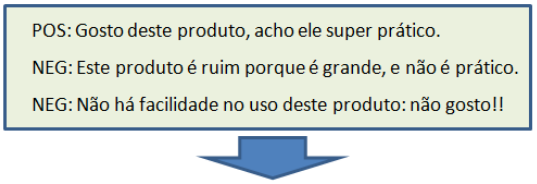


Figura 1.13. Aprendizado Supervisionado (Fonte: adaptado de [Tan et al. 2006])

contendo opiniões sobre produtos é convertido em um conjunto de treino contendo representações binárias dos termos usados (i.e. existe ou não existe no texto) . A classe alvo é o atributo *polaridade*, para o qual os valores são fornecidos junto com cada sentença.



gosto	produto	ruim	grande	prático	não	facilidade	uso	polaridade
1	1	0	0	1	0	0	0	pos
0	1	1	1	1	1	0	0	neg
1	1	0	0	0	1	1	1	neg

Figura 1.14. Conjunto de Treino : representação binária de termos

A qualidade do modelo preditivo é medida utilizando dados distintos daqueles usados como conjunto de treino (*dados de teste*). Podem ser usados dois conjuntos de dados distintos (treino e teste), ou métodos de validação cruzada, nos quais parte dos dados são usadas para treino, e parte para teste. No método *hold out* (um de fora), uma parte dos dados é separada para teste, tipicamente um terço. No método de validação cruzada *k-fold* (*k-fold cross validation)*, os dados são divididos em k partições de mesmo tamanho, onde frequentemente k=10 [Tan et al. 2006]. A partir disto, um subconjunto é utilizado para teste e os k-1 restantes são utilizados para treino, calculando-se a capacidade de generalização do modelo. Este processo é realizado k vezes, alternando de forma circular o subconjunto de teste, e ao final, é calculada a média das execuções.

A qualidade do modelo preditivo resultante da etapa de aprendizagem é medida em termos de métricas como *acurácia* (capacidade do modelo de prever corretamente), *precisão* (número de instâncias previstas corretamente em uma dada classe), *revocação*

(número de instâncias de uma dada classe previstas na classe correta), e medida F (que combina precisão e revocação).

Existem vários tipos de algoritmos de classificação. Entre os mais citados na área de mineração de opiniões estão [Pang et al. 2002,Pang and Lee 2008,Liu 2012,Tsytsarau and Palpanas 2012]:

Naïve Bayes (NB) : este algoritmo probabilístico é simples e bastante eficiente na classificação de textos em geral. Ele é baseado na aplicação do Teorema de Bayes com a premissa de total independência entre variáveis. Por exemplo, para prever a classe "cardíaco", assume-se que as *features* "colesterol" e "alimentação" não possuem relação uma com a outra. O modelo resultante para cada classe é a probabilidade dos valores assumidos por cada *feature*. No exemplo, é possível que a probabilidade de colesterol=alto seja alta, e alimentacao=saudável seja baixa. Este tipo de algoritmo trabalha bem tanto com *features* numéricas e discretas, e com alta dimensionalidade (i.e. muitas *features*).

Support Vector Machine (SVM) : um algoritmo de aprendizado linear que pode ser utilizado tanto para classificação, quanto para regressão. Um modelo SVM é uma representação dos exemplos do conjuntos de treino como pontos no espaço, mapeados de tal forma que os dados de cada categoria estejam separados pelo maior espaço possível entre eles. Mais precisamente, o SVM constrói um hiperplano que divide o espaço em duas dimensões (ou conjunto de hiperplanos, no caso de mais de duas dimensões). A separação entre as categoria é feita pelo hiperplano que maximiza a distância entre os pontos mais próximos entre quaisquer classes.

Maximum Entropy (ME) : outro classificador probabilístico que pertence à categoria dos modelos exponenciais. Contudo, ao contrário de NB, ele não assume a independência das *features*. Isto é interessante no caso de textos, porque nem todas palavras são independentes umas de outras. ME é baseado no princípio da Entropia Máxima, e dados de treino são usados como um conjunto de restrições sobre distribuições probabilísticas condicionais. ME dá preferência aos modelos de distribuição uniformes, que também satisfaçam restrições. Em outras palavras, a distribuição é uniforme em todas as classes, exceto em casos específicos expressos através dos dados de treino. Por exemplo, assuma um problema de 4 classes categorizando notícias, no qual o termo "prisão" tem 40% de chance de ser utilizado em notícias da classe "policial". Isto faz com que haja 20% de probabilidade de que o documento pertença a cada uma das outras 3 classes. Já se um documento não contiver o termo "prisão", a probabilidade é uniforme nas 4 classes (i.e. 25% para cada classe). Uma das desvantagens é que este algoritmo não tem desempenho tão bom quanto NB, devido à necessidade de otimização necessária para estimar os parâmetros do modelo. Por outro lado, ele tem se mostrado eficiente para problemas não binários, i.e. com mais de duas classes.

Redes neurais : Refere-se a uma família de algoritmos que se caracterizam por uma representação de nodos conectados por pesos, os quais são ajustados por algoritmo de aprendizado à medida que ajusta aos dados de treino. Assim, são capazes de

aproximar funções não lineares sobre as entradas (i.e. *features*) do modelo. Conceitualmente, os pesos adaptativos representam a força de conexão entre os neurônios, que são ativados durante treino e predição. As redes neurais se organizam de acordo com modelos específicos, como o SOM (Self-Organizing Maps) ou Multi-Layer Perpectron (MLP).

Todos estes algoritmos estão disponíveis em ambientes de mineração de dados populares, como o Weka[16], R[17] ou Rapid Miner[18]. Além disto, estes ambientes oferecem algumas facilidades para pré-processamento dos textos, visando preparar os dados para a classificação (e.g. tokenização básica, stemmer, remoção de stopwords, etc).

1.5.2.2. Preparação de *Features*

Apesar dos algoritmos citados na seção anterior terem sido usados com sucesso na mineração de opiniões, os resultados relatados são bastante influenciados pela preparação e seleção das *features* no conjunto de treino. Uma discussão completa está fora do escopo deste trabalho, e pode ser encontrada em obras clássicas sobre mineração de dados [Tan et al. 2006]. O enfoque será de discutir as preparações de *features* mais empregadas na mineração de opiniões, a saber:

- *palavras de sentimento:* uma escolha que deve ser feita é sobre os termos de um texto a serem utilizados como *features* . Alguns trabalhos buscam limitar os termos àqueles que expressam sentimento. Neste sentido, adjetivos e advérbios são os candidatos mais prováveis [Turney 2002], mas verbos e substantivos também são usados para expressar sentimentos [Liu 2012]. O exemplo da Figura 1.14 utiliza esta estratégia. Existem trabalhos que relatam melhores resultados quando todos os tipos de termos são usados sem distinção, considerando apenas remoção de *stop words*. A representação de termos de sentimento pode ser complementada com contagens de termos de sentimento com a ajuda de um léxico (e.g. número de termos positivos e negativos [Mohammad et al. 2013]).

- *representação binária ou pesos:* a representação binária consiste em informar se um termo aparece ou não no texto, como no exemplo da Figura 1.14. Esta representação é muito usada para textos curtos, como *tweets*, posts, ou análise em nível de sentença. A intuição por trás desta representação é que uma única ocorrência do termo é suficiente para determinar a polaridade da opinião. Por exemplo, ao dizer "Adorei a comida e adorei o ambiente", fica clara a opinião independentemente do número de vezes que o termo "adorar" foi usado. Em nível de documento, outra opção é utilizar um peso, como por exemplo, a frequência das palavras no texto, possivelmente normalizada pelo tamanho do documento, ou pela frequência do termo no conjunto dos documentos. A intuição neste caso é que, se em um texto mais longo o mesmo termo foi usado múltiplas vezes, ele tem um peso diferente

[16] www.cs.waikato.ac.nz/ml/weka
[17] www.r-project.org
[18] rapidminer.com

de outros utilizados menos frequentemente. Por exemplo, se em uma revisão de filme o termo “adorar” aparecer 3 vezes (e.g. para roteiro, o diretor e o figurino), e “detestar” uma vez (e.g. para um ator), existe um peso maior para o termo de conotação positiva. Outra opção para representação de pesos é o TDF-IF, bastante usado em recuperação de informações, que neste caso, dá importância a palavras mais raras no conjunto de documentos.

- *normalização e POS:* escolha pelo uso dos radicais ou dos lemas, ao invés dos termos originais, buscando uma unificação das diferentes variações dos termos em documentos distintos. Como já mencionado, o uso do radical nem sempre traz bons resultados, por eliminar as variantes que alteram a polaridade de termos com um mesmo radical. Uma forma de distinguir entre termos semelhantes é anexar o POS a cada termo. Por exemplo, diferenciar como “toler_N” e “toler_Adj” os termos “tolerância”e “tolerável” que têm o mesmo radical; ou diferenciar palavras com o mesmo lema, mas distintas classes gramaticais. Uma forma frequente utilizada para distinguir textos com sentimento e neutros é a contagem do número de termos de cada categoria (e.g. número de adjetivos, número de verbos, etc).

- *negação:* dependendo da estratégia de identificação de negação, a informação de negação pode ser anexada às *features* . Por exemplo, considerando o exemplo da Figura 1.14, ao invés de incluir o termo de negação “não” como *feature*, o termo “prático” seria representado por duas *features* : pratico e pratico_NEG. O mesmo para o termo “gosto”.

- *emoticons e outras formas de emoção:* pode-se representar a presença ou contagem de emoticons ou outras símbolos típicos da internet (e.g. LOL, kkkkk, rsrsrs), ou de expressões alongadas (e.g. "ameeeeei", "horrooooooor") que representam intensidade.

- *termos isolados e n-gramas:* uma ampla discussão existe neste quesito, com resultados contraditórios. O uso de n-gramas (n>1) visa capturar termos adjacentes (ou quase adjacentes) que caracterizam emoção (e.g. locuções, expressões idiomáticas, sentimentos contextuais).

Além de todas estas variações na preparação de *features*, o processo pode incluir ainda uma estrátégia para a busca das melhores *features* na construção do modelo (*feature selection*). Esta seleção é feita com a premissa de que as técnicas de classificação podem ser atrapalhadas por *features* redundantes ou irrelevantes. Técnicas conhecidas são Information Gain, Information Ration, Chi Squared, algoritmos genéticos, entre outras [Tan et al. 2006].

O atributo alvo é a mensuração da polaridade adotada. A classificação binária de opiniões (positiva ou negativa) permite atingir melhores resultados, e pressupõe que textos neutros foram eliminados na etapa de identificação. A adoção de um número maior de classes implica um cuidado extenso com o conjunto de treino, no tocante a número de exemplos para cada classe, e balanceamento dos exemplos entre as diferentes categorias.

1.5.2.3. Discussão

A classificação de polaridade usando o aprendizado supervisionado é bastante sensível à escolha das *features*, como discutido na seção precedente. Logo, resultados relatados na literatura sobre os melhores algoritmos e os pré-processamentos mais efetivos são bastante contraditórios. Uma das grandes limitações no uso de aprendizado supervisionado para definição de polaridade é a necessidade de dados rotulados para treino. O desempenho destes métodos é afetado não somente pela quantidade, mas igualmente pela qualidade dos dados de treino disponíveis. Ainda, cada conjunto de treino é fortemente vinculado ao seu domínio. Nos trabalhos envolvendo revisões de produto, a classificação dada pelos usuários na forma de notas ou estrelas é utilizada como classe do texto correspondente [Pang et al. 2002, Turney 2002]. Tal facilidade não está disponível para outros domínios, implicando a necessidade de anotações manuais, as quais são trabalhosas e com alto teor de subjetividade.

Existem algumas propostas para vencer a limitação da anotação dos dados. Em [Wiebe and Riloff 2005] foi proposto um método de classificação incremental baseado em regras, que utiliza regras genéricas sobre estruturas sintáticas para gerar texto subjetivo rotulado, o qual é usado para treinar classificadores e gerar novas regras mais específicas. A geração automática de texto rotulado no domínio de política também é proposta em [Sarmento et al. 2009]. Estas alternativas são contudo impraticáveis para fontes que geram dados em *streaming*, onde além de grandes volumes de dados que devem ser analisados com baixa latência, existe uma volatilidade enorme nos tópicos e termos utilizados. Para *tweets*, foi proposta a anotação automática com base na presença de emoticons. Esta abordagem foi avaliada em diferentes línguas, sendo que os piores resultados são para a língua portuguesa, que utiliza outro tipo de expressão de sentimento (e.g. "kkkkk"). Outra alternativa é uma abordagem semi-supervisionada voltada à análise de sentimentos em tempo real [Calais Guerra et al. 2011], que determina a polaridade de sentimento de *tweets* tomando como ponto de partida opositores e apoiadores conhecidos, e inferindo através de regras de associação a propagação deste sentimento nas conexões sociais.

1.5.3. Abordagens Estatísticas e Semânticas

As abordagens estatísticas, por vezes denominadas *não supervisionadas* [Liu 2010, Liu 2012], baseiam-se na premissa de que palavras que traduzem opiniões frequentemente são encontradas juntas no corpus dos textos. Se a palavra ocorre mais frequentemente junto a palavras positivas (negativas) no mesmo contexto, então é provável que seja positiva (negativa); já se ocorre em igual frequência, a palavra deve ser neutra. A polaridade de uma palavra desconhecida pode ser determinada calculando a co-ocorrência com uma palavra notadamente positiva (negativa), tal como "excelente" ou "péssimo". A técnica mais representativa nesta categoria é a Pointwise Mutual Information (PMI) [Turney 2002].

O PMI entre dois termos quaisquer x e y é calculado segundo a Equação 2, onde Pr(x e y) é a probabilidade de co-ocorrência dos termos x e y, enquanto que Pr(x).Pr(y) é a probabilidade de co-ocorrência se são estatisticamente independentes. Esta razão é portanto a medida de grau de independência estatística entre os dois termos, e o logaritmo desta razão é a quantidade de informação ganha se os termos são observados juntos. A

polaridade do sentimento de um termo x, dada pela Equação 3, é a diferença entre os valores de PMI calculados a partir de duas listas opostas: termos positivos (e.g. excelente), e termos negativos (e.g. péssimo). Na proposta original deste método [Turney 2002], as co-ocorrências foram obtidas a partir da internet, utilizando uma mecanismo de busca existente na época (AltaVista).

$$PMI(x,y) = \log_2\left(\frac{Pr(x \wedge y)}{Pr(x)Pr(y)}\right) \tag{2}$$

$$PMI\text{-}IR\ (x) = \sum_{p \in pWords} PMI(x,p) - \sum_{n \in nWords} PMI(x,n) \tag{3}$$

Outras técnicas na mesma abordagem são o Semantic-orientation Latent Semantic Analysis (SO-LSA) [Turney 2002], que usa a LSA para calcular a força da associação semântica entre dois termos, através da análise estatística entre eles; e a Latent Dirichlet Allocation (LDA), muito utilizada para a extração de tópicos em textos [Aggarwal and Zhai 2012]. A abordagem estatística é menos suscetível à dependência de contexto quando comparada às abordagens por dicionário e por aprendizado de máquina, e pode ser utilizada para complementá-las [Tsytsarau and Palpanas 2012]. Contudo, ela é dependente de um corpus de tamanho significativo.

A abordagem *Semântica* é bastante parecida com a estatística, exceto que a polaridade é calculada em termos de alguma medida de distância entre termos. O princípio das técnicas nesta categoria (e.g. [Hu and Liu 2004, Godbole et al. 2007]) é que palavras semanticamente próximas devem ter a mesma polaridade. Por exemplo, o WordNet [Fellbaum 2010] provê diferentes relacionamentos entre palavras que podem ser usadas para calcular a polaridade do sentimento, tais como sinônimos e antônimos. De forma similar à abordagem estatística, palavras semente com sentimento positivo e negativo são utilizadas como ponto de partida de um processo de comparação ou expansão, que busca determinar a polaridade dos termos. Isto pode ser feito através da mensuração da distância entre palavras usando as relações como caminho (e.g. [Godbole et al. 2007]), ou da contagem da frequência com que são associadas a sinônimos positivos/negativos (e.g. [Hu and Liu 2004]). A abordagem semântica também pode ser usada como complemento às outras abordagens, como forma de expansão ou de aquisição de vocabulário específico, na ausência de bons dicionários de sentimento. No entanto, ela carece ainda de métodos outros que manual, para validação da polaridade atribuída aos termos.

1.6. Estudo de Caso: Mineração de Opiniões em Comentários sobre Notícias Políticas

Esta seção ilustra como a mineração de opiniões foi aplicada em um estudo de caso real envolvendo comentários de leitores da versão on-line da Folha de São Paulo (Folha.com). Mais especificamente, visa-se através da mineração de sentimentos detectar a opinião de leitores sobre candidatos a eleições. Os comentários selecionados para análise são relacionados a notícias políticas envolvendo os candidatos. Três eleições foram monitoradas: presidenciais (2010), governo do estado de São Paulo (2010) e cidade de São Paulo (2012).

Este estudo de caso faz parte de uma pesquisa mais abrangente que visa utilizar métricas de sentimentos para prever variações na intenção de votos [Tumitan and Becker 2013, Tumitan and Becker 2014]. O framework de análise está descrito na Figura 1.15, e apresenta dois grandes componentes: *Mineração de Opiniões* e *Previsão de Variação de Intenção de Votos*. O componente *Mineração de Opiniões* tem por objetivo derivar séries temporais diárias que totalizam o sentimento positivo e negativo sobre cada candidato monitorado. Este sentimento é calculado totalizando menções positivas e negativas a cada candidato expressas em nível de sentença. As séries temporais de sentimento são utilizadas como entrada para o componente *Previsão de Variação de Intenção de Votos*, junto com uma série temporal mais esparsa representando as intenções de voto divulgadas por entidades de pesquisa (e.g. Ibope, Datafolha). Diferentes *features* são derivadas das séries temporais de sentimento, as quais são utilizadas para prever variações na intenção de voto (i.e. aumento, redução, inalterada). Maiores detalhes são descritos em [Tumitan and Becker 2014].

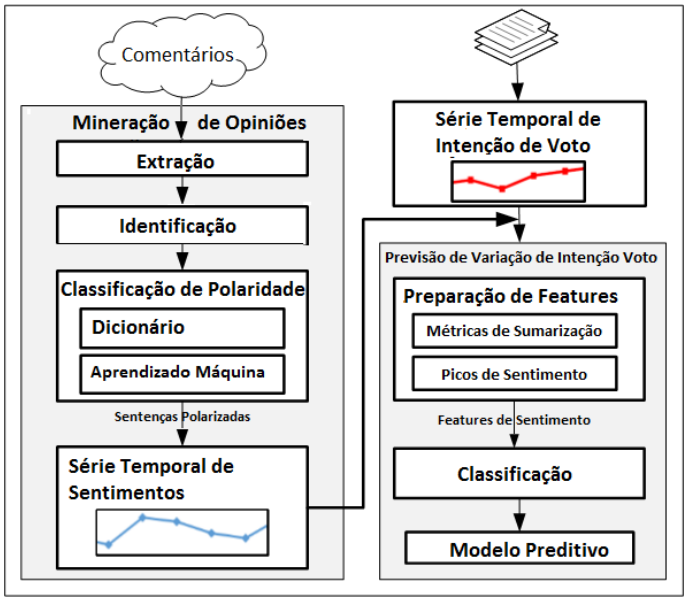


Figura 1.15. Framework de Previsão de Variações em Intenções de Votos baseada em Sentimentos (Fonte: [Tumitan and Becker 2014]).

Nesta seção nos limitaremos a discutir os aspectos mais relevantes da mineração de opiniões, ilustrando as etapas de Identificação de sentenças e seus alvos, e da Classificação da polaridade, na qual comparamos as abordagens baseada em léxico, e baseada em aprendizado supervisionado. Embora as diferentes etapas do processo sejam discutidas de forma linear, é importante destacar que a mineração de opiniões, assim como qualquer processo de descoberta de conhecimento [Fayyad et al. 1996], é um processo não-trivial, iterativo e interativo. Em função de resultados e percepções obtidas a cada etapa, passos anteriores são refeitos buscando aperfeiçoar os resultados. Esta seção conclui com uma breve discussão sobre como os resultados da mineração de opiniões foram agregados para diferentes propósitos.

1.6.1. Extração de Comentários

O conjunto de dados é composto de comentários gerados por leitores da Folha on-line a respeito de notícias sobre eleições. Para identificar as notícias relevantes, e extrair seus respectivos comentários, utilizou-se o Google Reader como indexador da seção política da Folha.com (tag Poder). O conjunto de dados contém notícias e seus respectivos comentários referentes a três eleições, e os dados coletados abrangem o mês que precedeu a data de cada eleição (primeiro turno). Para cada eleição, foram selecionados os candidatos com as maiores intenções de voto (e portanto, os mais comentados), a saber:

- Eleição Governamental de São Paulo (2010) : Aloízio Mercadante (PT) e Geraldo Alckmin (PSDB);
- Eleição Presidencial (2010) : Dilma Rousseff (PT), José Serra (PSDB) e Marina Silva (PV);
- Eleição Municipal de São Paulo (2012) : Celso Russomanno (DEM), Fernando Hadddad (PT) e José Serra (PSDB).

Utilizou-se o mesmo conjunto de comentários para analisar o sentimento das eleições governamentais e presidenciais de 2010, porque noticias e comentários estavam bastante atrelados aos partidos, fazendo associações dos candidatos dos diferentes níveis de eleições. Para distinguir entre estas duas eleições, utilizou-se o alvo da opinião. Por exemplo, uma menção a Dilma refere-se a eleição presidencial, e a Mercadante, à eleição governamental.

1.6.2. Análise Preliminar dos Dados

Um dos primeiros passos em qualquer processo de descoberta de conhecimento é uma análise superficial dos dados, visando estabelecer um profiling e adquirir uma familiaridade com as características dos dados que guiará as demais etapas do processo. A Tabela 1.2 resume algumas características dos conjuntos de dados. Observou-se uma variação bastante grande no número de comentários por notícia, bem como no tamanho dos comentários.

Uma amostra dos comentários é apresentada na Figura 1.16. Através de uma inspeção superficial manual sobre comentários aleatoriamente selecionados, detectamos alguns problemas importantes que deveriam ser considerados em nosso processo de mineração de opiniões:

- Múltiplos alvo dos comentários: observamos que um número expressivo de comentários envolviam opiniões sobre vários candidatos (e.g. Comentário 1 da Figura 1.16). Esta característica motivou a escolha pela análise da opinião em nível de sentença;
- Variações nas menções aos candidatos, tais como José Serra, Serra ou Zé Serra. Algumas das menções inclusive denotam sentimento (e.g. Vampisserra, Dilmais, Mhaldad). Esta característica requer reconhecimento das diferentes menções, bem como do sentimento implicado;

Tabela 1.2. Perfil dos dados da base de dados do primeiro turno das eleições analisada.

	Eleição de 2010		Eleição de 2012	
	Bruto	**Pré-processado**	**Bruto**	**Pré-processado**
Número de notícias	2.232	1.763	583	340
Número de comentários	225.217	190.975	36.108	25.115
Média de comentários por notícia (DP)	98.59 ($\pm$235.6)	86.09 ($\pm$206.5)	61.93 ($\pm$142.5)	44.06 ($\pm$92.5)
Número de sentenças	-	673.146	-	79.752
Média de sentenças por comentário (DP)	-	3.05 ($\pm$2.06)	-	3.17 ($\pm$2.34)
Comentários com menos de 4 palavras	5.148	0	7.185	0
Comentários com similaridade igual ou maior que 85%	29.094	0	3.808	0
Período	01/09/2010 até 03/10/2010		01/09/2012 até 07/10/2012	
Entidades	5 candidatos		3 candidatos	

- Termos específicos de sentimento: Observamos o uso de palavras com caráter de sentimento específicos (e.g. malufista, petralha, mensaleiro, como no Comentário 2), expressões idiomáticas e gírias ("é o cara"), ou internetês. Esta característica requer lidar com a habilidade de tratar com vocabulário específico ao contexto;
- Caracteres especiais para disfarçar palavras de baixo calão (e.g. f/d/p), possivelmente devidos à moderação do jornal (e.g. Comentário 3). Esta característica pode ser tratada na tokenização, visando reconhecimento e normalização das palavras;
- Erros no emprego da linguagem, incluindo ortografia, acentuação e gramática. Por vezes os erros são tão sérios, que impedem a compreensão do comentário (e.g. Comentário 5). Esta característica por um lado implica a necessidade de normalização dos termos (e.g. ladrão, ladrao, ladrãu). Por outro lado, reforçou a decisão de não empregar recursos analíticos avançados, que não oferecem bons resultados nestas situações. Outra opção, que não foi priorizada, foi a tentativa de correção de erros simples (e.g. grafia e acentuação);
- Elevado número de comentários irônicos (e.g. Comentário 4). Devido à complexidade desta questão, a ironia não foi tratada neste estudo de caso, mas seus efeitos fizeram-se sentir na baixa precisão da classificação de comentários positivos (i.e. elevado número de falsos positivos);
- Erros de manipulação dos usuários: foi observado um grande número de comentários em branco, ou com um alguns caracteres digitados aleatoriamente. Também detectamos a presença de comentários duplicados, ou quase duplicados. Esta característica pode ser devida a erros de manipulação no sistema, ou *spams*. Decidimos tratar estes comentários como ruídos, e removê-los do dataset para evitar viés.

1.6.3. Identificação de Sentenças com Menções

Uma vez obtidos os comentários e analisadas suas propriedades, iniciou-se a etapa de identificação de sentenças com menções às entidades monitoradas. O nível de análise es-

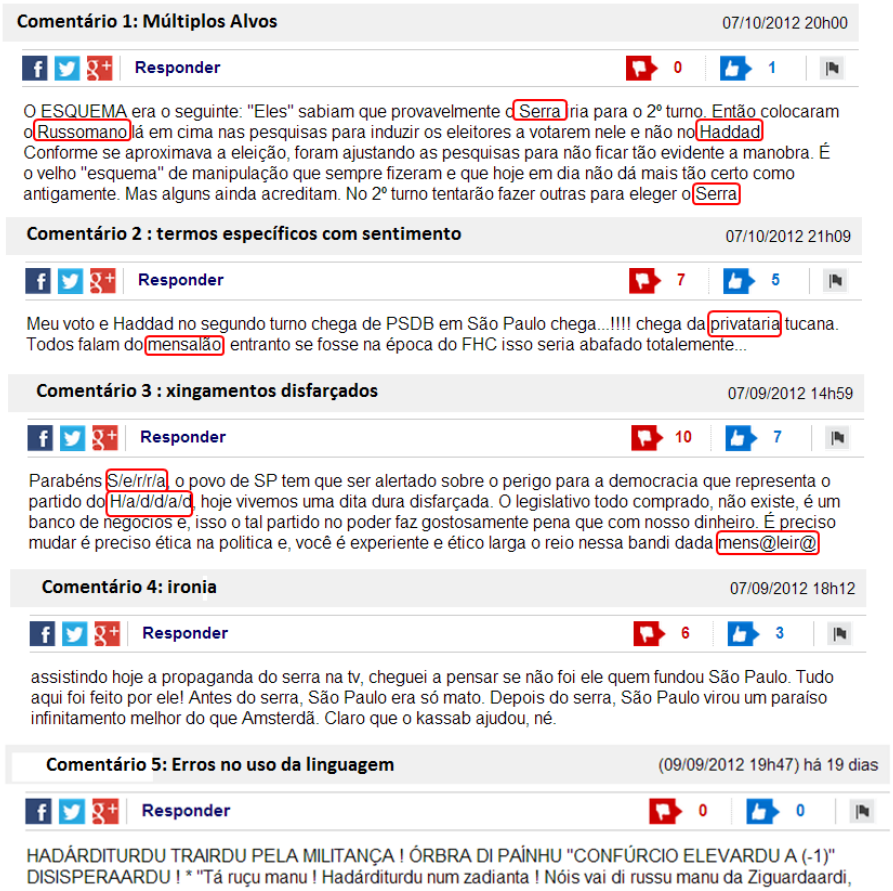


Figura 1.16. Exemplos de comentários

colhido foi o de sentença, porque um volume significativo de comentários envolvia múltiplas menções. Por exemplo, no Comentário 1 da Figura 1.16, existem duas sentenças com comentários sobre o candidatos Serra, e uma envolvendo dois outros candidatos (Haddad e Russomano). O processamento aplicado envolveu identificar comentários relevantes, quebrá-los em sentenças, e selecionar aquelas com menções. Um subproduto deste processo foi a aquisição de um léxico especializado. As principais decisões são discutidas nesta seção. Ressalta-se novamente que, embora atividades e decisões apresentadas sequencialmente e de forma linear, o processo foi altamente iterativo e interativo.

1.6.3.1. Eliminação de Ruídos

Foram eliminados todos os comentários considerados ruídos, isto é, comentários muito curtos (menos de 4 palavras) ou duplicados. Para determinar os comentários (quase) duplicados, foram feitos experimentos utilizando duas funções de similaridade, Jacquard e Coseno, as quais são bastante utilizadas para definição de similaridade em documentos. Utilizou-se os resultados obtidos através da função Coseno, com um limiar de corte de 85% [Tumitan and Becker 2013]. A quantidade de comentários duplicados foi significativa (12% e 10% nos datasets de 2010 e 2012, respectivamente), reforçando nossa decisão de remove-los do dataset.

1.6.3.2. Tokenização e Outros Tratamentos de PLN

Para a tokenização, foram escolhidos recursos para quebra em sentenças e *tokens* disponíveis no NLTK (Package Tokenizer). Para quebrar em sentenças, foi utilizado o *punkt*, que possui um corpus de treinamento em português. Para identificação de *tokens*, foram utilizadas estratégias de quebra por espaços em branco, bem como por expressões regulares. Também foram empregados recursos do Unix para PLN (Unix for Poets) e funções em Python para problemas específicos.

Antes da quebra em sentença e tokenização, os comentários foram pré-processados para tratamento de caracteres específicos da Web, como marcações HTML/XML, tratamento de caracteres HTML (e.g. conversão de   para espaço em branco), identificação de informação relevante (e.g. autor e timestamp), entre outros.

Como parte da tokenização, foram tratadas as seguintes questões:

- Conversão para caracteres minúsculos;
- Remoção de caracteres de acentuação e cedilha;
- Tratamento de expressões disfarçadas (e.g. palavrões) usando heurísticas. Por exemplo, o caracter @ foi convertido no caractere a.

Foram feitos experimentos com a extração de radicais, utilizando RSLP, mas os resultados foram bastante insatisfatórios quando da classificação de polaridade das sentenças. Vários outros pré-processamentos de texto discutidos na Seção 1.3 não puderam ser aplicados, quer pela falta de recursos voltados à língua portuguesa do Brasil, quer pelas características dos comentários gerados por usuários. Por exemplo, não foram aplicados nem a lematização, nem o etiquetamento POS devido à falta de recursos. Fizemos experimentos de análise sintática nas sentenças utilizando o software Palavras, mas este não apresentou bons resultados devido aos erros de português (estruturais, sintáticos e de ortografia), e emprego de termos informais. Especificamente, tentamos quebrar sentenças em cláusulas para tratar a negação, bem como identificar mais precisamente o alvo de palavras de sentimento. No tratamento de negação, ficamos limitados ao emprego de janelas de distância entre termos de negação e palavras de sentimento, mas esta técnica não apresentou bons resultados na classificação da polaridade.

1.6.4. Identificação de Menções

Menções aos candidatos foram identificadas utilizando uma lista de referências conhecidas (e.g. Serra, Jose Serra, Ze Serra, Zeh Serra). Além disto, utilizando expressões regulares, buscou-se variações sobre os nomes dos candidatos. Por exemplo, uma expressão como *Serra* permite buscar menções como "Vampiserra" ou "QueSerraDeNós". As expressões resultantes foram verificadas manualmente, e quando implicavam sentimento, incluídas em um léxico específico. Como resultado destas ações foram filtradas 80.469 sentenças com menções, sendo 69.490 relativas aos candidatos de 2010, e 10.979 relativas a 2012.

É importante ressaltar que estas heurísticas implicam limitações que comprometem a correta identificação de menções. Uma delas é que menções que não envolvam

variações sobre o nome de um candidato não serão identificadas. Por exemplo, um dos candidatos era mencionado usando a expressão “Mr. Burns”, que ressaltava sua similaridade com um personagem de cartoon. Outra limitação importante é que não foi possível tratar a co-referência, novamente pela falta de recursos. Assim, é possível que muitas sentenças com opiniões tenham sido descartadas neste processo, pela incapacidade de conectar pronomes com as respectivas menções feitas em sentenças distintas. Por exemplo, no comentário “Vou votar em X. Ele é o melhor candidato”, após a quebra de sentenças, somente a primeira seria considerada para polarização. Finalmente, não foram tratadas adequadamente sentenças com mais de uma menção. Neste caso, assume-se que a polaridade da sentença será atribuída a cada candidato identificado. Por exemplo, na frase “X é tão ladrão quanto Y”, a polaridade negativa será atribuída a X e a Y. Contudo, na frase “X é muito melhor que Y” isto também ocorrerá, ainda que a opinião sobre X e Y seja diferente.

1.6.5. Classificação da Polaridade

Considerando as sentenças com menções extraídas na etapa anterior, foram testados dois métodos de classificação de polaridade. O método baseado em Dicionário tem a vantagem de ser simples, e não necessitar de um conjunto de treino. Contudo, esbarramos na ausência de bons léxicos de sentimento para a língua portuguesa, e no emprego de vocabulário específico não apenas voltado ao contexto das eleições, mas aos termos específicos utilizados para cada candidato. No método baseado em aprendizagem de máquina, a grande dificuldade é a necessidade de um corpus de treinamento. Para ambas abordagens foi necessária a criação de um corpus de referência através da anotação manual de sentenças que no mínimo permitisse avaliar o desempenho dos classificadores. Estes aspectos são discutidos no restante desta seção.

1.6.5.1. Corpus de Referência

Foi construído um corpus de referência para avaliar o desempenho da classificação de sentimento baseada em dicionário, e para usar como corpus de treinamento no método baseado em aprendizado de máquina. Com estes propósitos, foram selecionadas aleatoriamente 1.000 e 600 sentenças da base de dados das eleições de 2010 e 2012, respectivamente. Para cada conjunto de dados foram utilizados 3 anotadores com graduação em ciência da computação, e sem prévia experiência de anotação de corpus. Eles foram instruídos a basear suas anotações no conteúdo que estava explicitamente escrito, desconsiderando qualquer tipo de suposição a respeito de políticos ou partidos políticos, para que suas visões políticas não interfiram em seus julgamento [Sarmento et al. 2009]. O nível de concordância entre anotadores está indicado na Tabela 1.3, e pode-se notar que é bastante baixo. Utilizamos em nossos experimentos sentenças com pelo menos duas concordâncias, representando 92,7% das sentenças anotadas para 2010, e 97% para as eleições de 2012.

Para as eleições de 2010, foram obtidas 356 sentenças anotadas como negativas, 154 como positivas, e 417 como neutras. O corpus de referência para as eleições de 2012 contém 482 sentenças anotadas como negativas, 72 como positivas, e 28 como neutras.

Tabela 1.3. Concordância entre anotadores das eleições de 2010 e 2012.

Eleição	**2010**			**2012**		
Anotadores	**A vs. B**	**B vs. C**	**C vs. A**	**A vs. B**	**B vs. C**	**C vs. A**
Porcentagem	59.60%	49.40%	48.90%	75.66%	73.33%	74%
Todos concordam	33.10%			63%		
Todos discordam	7.30%			3%		

Embora as instruções para as anotações dos dois conjuntos de dados fossem as mesmas, notamos diferenças entre os resultados. Ressalta-se que apenas um anotador era comum nas duas anotações. Comparado com o de 2012, o conjunto de sentenças das eleições de 2010 revelaram uma proporção mais alta de sentenças anotadas como positivas, e um número significativo de sentenças neutras, o que pode ter influenciado os resultados. Durante o processo de anotação, também foram identificados termos correspondendo a expressões regionais e idiomáticas, apelidos, e palavras de sentimento informais, os quais foram incluídos no dicionário especializado. Notou-se que alguns termos de sentimento eram específico a cada candidato, e a cada eleição. Conclui-se com esta experiência que processo de anotação é trabalhoso, bastante subjetivo, e seus resultados devem ser usados com cautela.

1.6.5.2. Classificação baseada em Dicionário

Para a classificação baseada em dicionário, o melhor dicionário encontrado foi o Sentilex-PT, para português de Portugal. O método de classificaçã de polaridade é muito simples: a polaridade de cada termo é buscada no dicionário, sendo somadas as polaridades positivas (1) e negativas (-1) dos termos encontrados no léxico utilizado.

Os resultados desta abordagem em nossos *datasets* foram bastante limitados, em particular para as sentenças positivas. Usando o *dataset* das eleições 2012, fizemos vários experimentos tentando melhorar estes resultados. Cada abordagem foi ainda experimentada utilizando uma tentativa de tratamento da negação baseada em uma janela de distância entre o termo de sentimento, e um termo de negação. A Tabela 1.4 resume os principais resultados dos experimentos, usando as medidas de acurácia, precisão, revocação e medida-F. São consideradas apenas as classes *positiva* e *negativa*.

- Baseline: esta abordagem consiste no uso direto do Sentilex-PT, considerando apenas a tokenização básica das palavras e remoção de *stop words*.

- Léxico Especializado: observamos que o léxico Sentilex-PT não capturava corretamente termos do português do Brasil, bem como gírias, expressões idiomáticas ou termos dependentes do contexto. Assim, criamos um léxico especializado a partir das seguintes ações: a) uso das menções de sentimento citadas na Seção 1.6.4; b) identificação de termos e expressões durante o processo de anotação; c) seleção dos 1.000 termos mais frequentes que não foram encontrados no dicionário, seguida da análise para identificação daqueles que implicassem sentimentos.

Tabela 1.4. Acurácia (P), Precisão (P), Revocação (R) e Medida-F (F) para cada variação da classificação baseada em léxico.

Variação	Abordagem	A(%)	Polaridade	P (%)	R (%)	F(%)
Baseline	Com Negação	34.11	Positiva	18.35	25.64	21.39
			Negativa	89.53	35.48	50.82
	Sem Negação	35.54	Positiva	17.35	21.79	19.32
			Negativa	89.22	37.76	53.06
Dicionário	Com Negação	42.68	Positiva	24.87	62.82	35.64
			Negativa	89.62	39.42	54.76
Especializado	Sem Negaçao	43.21	Positiva	26.02	65.38	37.23
			Negativa	90.52	39.63	55.12
Sem	Com Negação	52.14	Positiva	26.99	56.41	36.51
			Negativa	89.86	51.45	65.44
Acentuação	**Sem Negação**	**52.14**	Positiva	26.99	56.41	36.51
			Negativa	90.18	51.45	65.52
Stemming	Com Negação	34.82	Positiva	21.28	38.46	27.40
			Negativa	89.19	34.23	49.48
	Sem Negação	37.32	Positiva	21.01	32.05	25.38
			Negativa	87.62	38.17	53.18

- Sem Acentuação: nesta abordagem, além do uso de léxico especializado, removemos todas as acentuações e cedilhas, tanto das sentenças quanto dos léxicos;
- Stemming: além do uso de léxico especializado e da remoção da acentuação, foi aplicado um stemmer sobre as sentenças.

Pode-se verificar que o desempenho de classificação para sentenças negativas é nitidamente superior (próximo a 90% de precisão), e todas as abordagens tentadas visaram melhorar a classificação das sentenças positivas. Contudo, o melhor resultado obtido para a classe positiva não ultrapassou 27%. A melhor acurácia foi apresentada pela abordagem com léxico especializado sem acentuação, com nenhuma diferença significativa em relação ao tratamento (ou não) da negação. Os piores problemas foram o emprego de ironia, e o emprego de termos positivos, mas que não se referiam ao alvo. Outro fator que contribuiu bastante para estes resultados foi a agregação de polaridades quando termos de diferentes polaridades são encontrados na mesma sentença (e.g. “O candidato Y é uma bela droga"). No método utilizado, os termos “bela"e “droga" se anulam quando suas polaridades são agregadas, resultando na classificação da sentença como neutra.

1.6.5.3. Classificação baseada em Aprendizado de Máquina

Em nossa pesquisa, havíamos inicialmente utilizado apenas o conjunto de dados de 2012, e as 600 sentenças anotadas visaram verificar o desempenho do método baseado em dicionário [Tumitan and Becker 2013]. Insatisfeitos com os resultados, decidimos experimentar um método baseado em aprendizado supervisionado, o que nos levou a adotar um

segundo conjunto de dados (2010), e executar um novo processo de anotação para gerar um conjunto de treino mais substancial.

Testamos diferentes tipos de preparações de *features*, dentre as citadas na Seção 1.5.2.2. Os melhores resultados foram obtidos com uso de: a) unigramas, b) representação por pesos utilizando TDF-IF, e c) uso de Information Gain como critério para seleção de *features* . Observamos que havia uma grande influência da menção do candidato na determinação da polaridade, particularmente para candidatos com maior nível de rejeição. Para eliminar este viés, eliminamos das sentenças todas as menções a candidatos na sua forma padronizada. Testamos ainda outras preparação de *features*, mas com resultados inferiores: a) bigramas e trigramas; b) representação binária dos termos; c) remoção de *stop words*, e d) outras funções de seleção de *features* . Não utilizamos stemmer pelo fraco resultado apresentado nos experimentos com a classificação baseada em léxico. Pelos problemas já citados de falta de recursos, não restringimos as *features* a palavras de sentimento, não exploramos POS, nem aplicamos lematização.

Todos estes experimentos foram realizados com diferentes classificadores disponíveis no ambiente Weka, entre eles Naïve Bayes, EM, SVM, baseado em regras, baseado em árvores de decisão. Treinamos todos os classificadores usando sentença positivas e negativas, com *10 fold cross-validation* como método de verificação de desempenho. Os melhores resultados foram obtidos com o emprego de SMO (Sequential Minimal Optimization), uma implementação da família SVM de classificadores. Finalmente, para verificar o ajuste excessivo do modelo aos dados (i.e. *overfitting*), ainda testamos o SMO com a melhor preparação de *features* encontrada, utilizando o conjunto de dados de 2010 como treino, e o de 2012 como dados de teste.

A Tabela 1.5 apresenta uma comparação dos melhores resultados obtidos com a classificação baseada em léxico, classificação com SVM e *10 fold cross-validation*, e classificação com SVM usando conjuntos de treino e teste distintos. Observa-se que os resultados de desempenho são distintos conforme o conjunto de treino usado, o que mostra a dependência desta abordagem com o processo de anotação. Também é possível observar a superioridade do método baseado em aprendizado de máquina. Observa-se que a classificação de sentenças positivas, de uma maneira geral, continua com resultados insatisfatórios.

Embora os resultados do SVM com validação cruzada sejam os melhores, entendemos que eles são devido a *overfitting*, e o desempenho da abordagem com dois conjuntos de dados distintos é provavelmente mais realista. Para investigar esta questão, aplicamos uma função de seleção de *features* baseada em *Information Gain* em ambos conjuntos de dados e treino. Com os atributos mais relevantes, pudemos verificar um vocabulário muito distinto utilizado para expressar opiniões em cada eleição. Por exemplo, no conjunto de 2010, houveram muitas alusões às primeiras candidatas à presidência, e palavras como “presidenta”, “primeira”, “história”, e “guerreira” foram usadas para expressar opiniões positivas. Já nas eleições de 2012, ambos candidatos haviam sido ministros, e seus feitos nas áreas de edução e da saúde eram utilizados, bem como alusões à corrupção nos respectivos partidos (e.g. “genérico”, “enem”, “mensalão”, “malufista”).

Tabela 1.5. Comparação das diferentes abordagens usando Acurácia (A), Precisão (P), Recall (R) e medida F (F).

Eleiçao	Abordagem	A(%)	Polaridade	P (%)	R (%)	F(%)
2010	Baseado em Dicionário	50.39	Positive	29.39	62.99	40.08
			Negative	54.79	44.94	49.38
	SVM	**81.37**	Positive	**70.63**	**65.58**	**68.01**
			Negative	**85.56**	**88.20**	**86.86**
2012	Baseado em Dicionário	52.14	Positive	26.99	**56.41**	**36.51**
			Negative	**90.18**	51.45	65.52
	SVM	**83.24**	Positive	**27.59**	10.53	15.24
			Negative	86.45	**95.38**	**90.70**
	SVM (dados de teste 2012)	77.40	Positive	25.56	30.26	27.71
			Negative	87.98	85.27	86.61

1.6.6. Sumarização

O resultado do processo de mineração de dados é um conjunto de sentenças polarizadas. A etapa final do processo é agregar as polarizações individuais em série temporais diárias por candidato, as quais servem aos seguintes propósitos: a) visualização da evolução de sentimento por candidato; e b) previsão de variação nas intenções de voto.

Por exemplo, a série temporal da Figura 1.17 ilustra o total de sentimento positivo e negativo para um determinado candidato no período observado. É possível notar que as linhas são relativamente espelhadas, mas com uma quantidade de sentimento negativo bem maior. Uma análise mais detalhada dos resultados nos permitiu resumir o seguinte comportamento dos leitores [Tumitan and Becker 2013]: a) de uma forma geral, os comentários são negativos, revelando uma frustração geral com a política; b) poucos são os comentários usados para apoiar candidatos; c) muitos dos comentários revelam uma transferência de opinião do partido/membro do partido para o candidato. Este é um comportamento bastante distinto do encontrado no *twitter* (e.g. [O'Connor et al. 2010,Calais Guerra et al. 2011, Tumasjan et al. 2010], no qual existe endosso ou contraponto a candidatos.

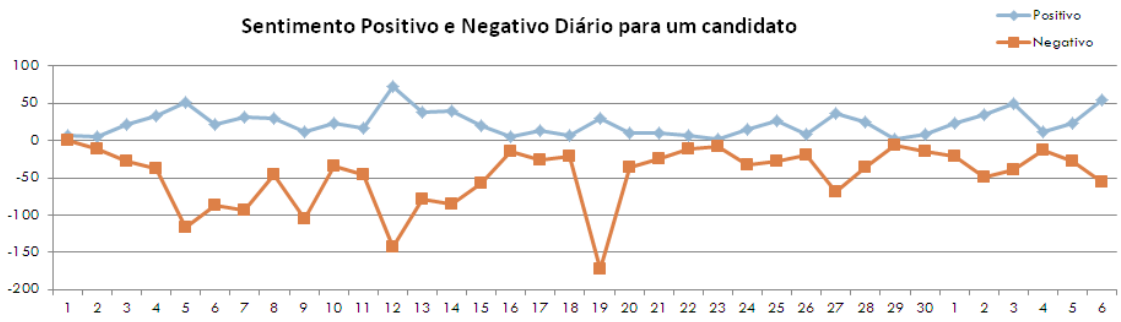


Figura 1.17. Exemplo de série temporal com sentimento para um candidato

Para a previsão de variação de intenções de votos, propusemos vários tipos de métricas, as quais podem ser agrupadas em duas categorias: totalização de sentimentos (e.g.

razão de sentimento positivo e negativo) e picos de sentimento (i.e. volume acima/abaixo do normal de sentimento positivo/negativo). Examinamos como derivar destas métricas diferentes tipos de *features* de sentimento, com as respectivas variações no acúmulo temporal de sentimento, e experimentamos estas preparações na previsão de variações de intenção de votos. Atingimos cerca de 80% de precisão na previsão, sendo que maiores detalhes podem ser encontrados em [Tumitan and Becker 2014].

1.7. A Mineração de Opiniões e suas Fontes de Dados

1.7.1. Revisão de Produtos

Uma parcela significativa dos trabalhos na área de mineração de opiniões concentra-se na revisão de produtos, como já mencionado. Por um lado, este foco é justificado pelo interesse comercial nesta classe de aplicação, e a pela grande disponibilidade de dados. Mas do ponto de vista computacional, revisões de produtos apresentam uma série de propriedades que facilitam a mineração de opiniões, quando comparadas a textos em geral, tais como predominância de conteúdo de opinião, bom volume de opinião sobre uma mesma entidade ou relativa a um domínio, e definição clara da entidade alvo. Ainda, o uso de vocabulário muito coloquial, gírias, *emoticons* e mesmo de ironia é bem mais limitado, se comparado a outros tipos de mídia (e.g. Twitter, comentários). Assim, dado o foco claro, e a menor quantidade de ruído, os problemas computacionais inerentes podem ser mais facilmente identificados, estruturados e tratados [Liu 2012]. Nesta seção, examinaremos trabalhos pioneiros nesta área [Turney 2002, Pang et al. 2002, Hu and Liu 2004]. Para uma compilação de avanços nesta área podem ser consultadas obras como [Tang et al. 2009, Liu 2012, Tsytsarau and Palpanas 2012].

1.7.1.1. *Analise de Opinião em Nível de Documento*

Um dos trabalhos precursores nesta área foi desenvolvido por Turney em [Turney 2002]. A motivação deste trabalho é agregar ao resultado de uma busca sobre revisões de um produto/serviço, informação sobre a recomendação (*Thumbs up*) ou não (*Thumbs down*) deste. Suas principais características são: a) análise em nível de documento, b) uso de abordagem estatística para classificação de polaridade, c) identificação de porções de texto com sentimento usando informação morfológica.

No tocante à etapa de identificação, assume-se que uma mecanismo de busca retorna uma coleção dos documentos referentes à entidade consultada, a qual é considerada como alvo de todas opiniões expressas nos documentos. A etapa de classificação envolve, para cada documento, seu pré-processamento para identificação *features* de interesse com base nas classes morfológicas das palavras, e a definição de sua polaridade. Finalmente, esta proposta não envolve explicitamente a etapa de sumarização, mas os documentos polarizados podem ser a entrada para várias aplicações, tais como geração de estatísticas, identificação de fóruns com posts abusivos (*flames*), etc.

Um documento de entrada é analisado usando um *parser* de POS, que rotula cada palavra do texto com sua classe gramatical. São então selecionados apenas pares de palavras que seguem determinados padrões sintáticos. Mais especificamente, retém-se pares de palavras onde um dos elementos é um adjetivo ou advérbio, e o outro, uma palavra

que lhe dá contexto. Com o contexto, a polaridade de palavras de sentimento pode ser melhor avaliada, tais como o adjetivo "inesperado", que possui conotação positiva no contexto de filme ("final inesperado"), ou negativo em uma revisão sobre carros ("defeito inesperado").

Para cada expressão resultante do pré-processamento, é calculado o PMI-IR, com base no PMI da expressão analisada com as palavras *excelente* e *ruim*, respectivamente. Para cálculo do PMI são submetidas consultas à internet com a mecanismo de busca AltaVista, onde o número de documentos retornados é utilizado como cálculo de frequência. O Altavista disponibilizava o operador *near*, que restringe a busca a documentos onde os termos estão próximos dentro de uma janela de no máximo 10 palavras. A polaridade final do documento, denominada *orientação semântica*, é dada pela média da polaridade de todas as expressões consideradas.

Como avaliação, foram usadas 410 revisões extraídas do *site* Epinions sobre filmes, automóveis, bancos e destinos de viagem, as quais estão associadas com uma classificação de estrelas. Foi obtida uma acurácia média de 74,39% em relação à recomendação do produto/serviço avaliado. É interessante notar que os piores resultados obtidos foram no domínio de filmes, já que mesmo que o sentimento geral sobre o filme tenha sido bom, os usuários sempre criticaram negativamente aspectos específicos. Ainda, não foi possível distinguir a opinião sobre o filme, do seu conteúdo objetivo. Por exemplo, na revisão *"Ele falava de um jeito devagar e metódico. Eu amei! Isto fez ele mais arrogante e mais perverso."*, as expressões "mais arrogante" e "mais perverso" referem-se ao personagem, mas são interpretadas como opiniões sobre o filme. Essa situação não aconteceu nos outros contextos analisados.

O desempenho devido à grande quantidade de buscas necessárias, e os restritos padrões gramaticais considerados, são apontados como os fatores mais limitantes desta proposta.

Uma alternativa à etapa de classificação foi apresentada na mesma época por Pang et al. em [Pang et al. 2002], usando métodos de aprendizagem de máquina. Mais especificamente, o trabalho concentrou-se em comparar experimentos com três classificadores conhecidos (SVM, NB e ME), e diferentes alternativas de pré-processamento. Na preparação de um vetor de termos, foram consideradas as seguintes alternativas de pré-processamento, algumas delas combinadas:

- unigramas com representação binária: adoção de qualquer termo (após remoção das *stop words*, e sem *stemming*). Quando uma negativa aparecia próxima ao termo, foi adicionado um prefixo NOT_ para representar a negação. Os termos mais frequentes na coleção (aproximadamente 18.000) foram representados como um vetor binário de termos. Experimentos foram feitos com e sem *stemming*;

- unigramas com peso: semelhante aos unigramas acimas, mas com uma representação vetorial com pesos (TDF-IF);

- bigramas: foram selecionados os bigramas mais frequentes, sem negação, os quais foram representados como um vetor binário;

- uso de POS: os termos foram rotulados usando um *parser*. Experimentos foram feitos considerando somente adjetivos como *features*, ou prefixando cada termo com classe gramatical para dar mais contexto;
- uso de posição: uso de unigramas com indicação de sua posição no texto (início, meio e fim). A intuição é que pessoas iniciam o texto, ou o concluem com a opinião mais importante, e que portanto a posição das palavras de opinião poderia ser importante.

Experimentos foram conduzidos sobre uma base de revisões sobre filmes, da qual foram extraídas 700 revisões positivas e 700 negativas. Em termo de qualidade, os resultados foram bastante semelhantes aos obtidos por Turney em [Turney 2002]. Os experimentos não revelaram de forma consistente a superioridade de nenhum dos classificadores adotados, e em termos de preparação de dados, o pré-processamento de vetor binário de unigramas foi, de uma forma geral, o mais eficaz.

1.7.1.2. *Análise de Opinião em Nível de Aspecto*

A análise de um produto em nível de documento permite derivar uma avaliação geral do mesmo. Em revisões onde existem sentimentos mistos expressos sobre a mesma entidade, pode-se chegar a uma situação de neutralidade em caso de médias (e.g. [Turney 2002]), onde as expressões positivas e negativas se anulam; ou de incapacidade de classificação correta, pela falta de *features* discriminantes (e.g. [Pang et al. 2002]). A análise em nível de aspecto permite detalhar o alvo do sentimento, de tal forma que possam ser detectados seus pontos fortes e fracos.

Um trabalho pioneiro nesta área foi proposto por Hu e Liu em [Hu and Liu 2004], que se caracteriza por: a) identificação de aspectos e de sentenças de opinião na etapa de identificação; b) uso de POS para identificação tanto de aspectos, quanto de palavras de sentimento; c) emprego da abordagem semântica para classificação da polaridade; e d) criação de sumários contendo o sentimento positivo e negativo sobre cada aspecto do produto.

A fase de identificação reconhece tópicos e sentenças de opinião. É primeiramente realizada a marcação das categorias gramaticais usando um *parser* de POS sobre os documentos. Para identificar os aspectos, são usadas regras de associação para encontrar termos frequentemente associados, os quais são podados segundo algumas regras de pré-processamento. Para encontrar as sentenças de opinião, são encontrados os adjetivos, considerados como palavras de opinião, e se estão próximos a aspectos, são designados como *opiniões efetivas*.

A etapa de classificação polariza as opiniões sobre os aspectos usando uma abordagem semântica, e refina o conjunto de tópicos utilizando as palavras de opinião. Para a polarização, o ponto de partida é um conjunto de palavras semente com polaridade definida manualmente. Usando o WordNet, a polaridade destas palavras é propagada para sinônimos (mesma polaridade) e antônimos (polaridade inversa), resultando assim em um dicionário específico de sentimentos para o domínio das revisões. Este algoritmo de propagação é iterativo, de tal forma que a polaridade antes desconhecida de uma determinada

palavra, acaba sendo encontrada em uma iteração futura, já que uma vez que um adjetivo é polarizado, ele passa a incorporar o conjunto de sementes. Para relacionar o sentimento ao aspecto, são usadas apenas as palavras de sentimento efetivas, na mesma sentença ou na sentença próxima.

As palavras de sentimento polarizadas são também utilizadas para identificar os aspectos de produtos menos frequentes e previamente desconhecidos. Por exemplo, se a palavra "excelente" é uma palavra de sentimento conhecida e na sentença, perto dessa palavra, existe uma locução nominal, assume-se que esta seja um aspecto de produto.

Finalmente, os sumários são criados em dois passos: 1) para cada aspecto, é feita uma contagem de quantas revisões opinam de maneira positiva/negativa; 2) aspectos são ordenados pela frequência com que aparecem nas revisões de produtos. Sentenças que contribuíram à pontuação positiva/negativa são mostradas junto com os sumários

A validação envolveu experimentos com vários aparelhos eletrônicos, usando revisões extraídas dos *sites* Amazon.com e Cnet.com. Estas foram manualmente anotadas quanto ao seus aspectos, sentenças opinativas e sua respectiva polaridade. Como resultado, obteve-se uma precisão média de 68,25% para aspectos identificados, de 64,2% para identificação de sentenças opinativas e uma acurácia média de 84,2% no tocante à polaridade. Os autores apontam como oportunidade de melhorias a resolução de pronomes, o uso de palavras de sentimento de classes outras que adjetivos, e o tratamento da intensidade da opinião sobre os aspectos extraídos.

1.7.2. Mineração de Opiniões em Notícias e Blogs

A mineração de opiniões em textos não estruturados é bem mais complexa que em revisões de produtos. Existem muitos desafios a serem enfrentados, entre eles: a) o texto pode conter opiniões sobre múltiplos alvos, e não é fácil reconhecê-los; b) o conteúdo de opinião é mais esparso no texto; c) dados os dois problemas anteriores, a associação da entidade alvo com a opinião fica ainda mais complexa; d) alguns tipos de texto, como notícias, tendem a não explicitar a opinião diretamente, fazendo-o através de artifícios (e.g. frases atribuídas a outras pessoas citadas na notícia); e e) é difícil distinguir entre conteúdo ruim (e.g. um terremoto), e uma opinião boa sobre um conteúdo ruim (e.g. elogiar o socorro às vítimas de um terremoto). Nesta seção são discutidos dois trabalhos que mineram opiniões sobre notícias: um que utiliza menções a entidades específicas [Godbole et al. 2007], e outro que aborda os desafios de encontrar sentimento explícito em notícias [Balahur et al. 2009b, Balahur et al. 2010].

1.7.2.1. *Monitoramento de entidades em jornais e blogs*

Godbole *et al.* em [Godbole et al. 2007] têm como objetivo desenvolver um sistema de análise de sentimentos em notícias e blogs que monitore o sentimento do público geral em relação a determinadas entidades, como pessoas, locais ou marcas. Assume-se que as entidades analisadas possuem características singulares, tais como atletas, celebridades, políticos, criminosos, etc, e que as opiniões emitidas devem ser interpretadas dentro de cada contexto. O método propõe uma forma algorítmica de construir dicionários de sentimentos voltados a cada contexto, e métricas para medir o sentimento expresso so-

bre essas entidades. As principais características desta proposta são: a) análise em nível de sentença; b) uso de menções específicas para identificação de entidades; c) uso de abordagem semântica de criação de léxicos voltados a cada domínio, os quais são usados na polarização; e d) representação do sentimento através de diferentes métricas, cuja evolução pode ser monitorada.

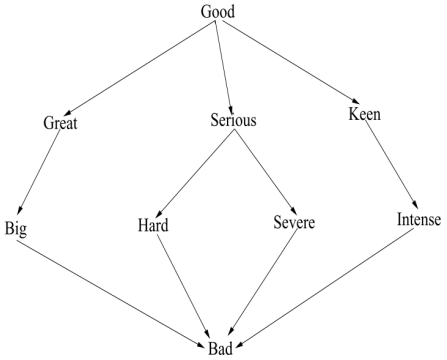


Figura 1.18. Expansão da palavra "*good*" até uma inversão de sentimento (Fonte: [Godbole et al. 2007]).

Na fase de identificação, são encontradas as menções às entidades de interesse, e as respectivas sentenças. O método de co-ocorrência é utilizado para encontrar variações nas menções à mesma entidade, e uma ferramenta é utilizada para resolver pronomes. Usando os léxicos específicos relativos ao domínio da entidade em questão, cada sentença é polarizada. Se a entidade e uma palavra do léxico estiverem na mesma sentença, sua polaridade é atribuído ao alvo.

Para a fase de sumarização, são propostas métricas de sentimento, envolvendo *polaridade* e *subjetividade*. A primeira busca determinar a razão entre sentimento positivo e negativo, enquanto que a segunda, a razão entre sentimento positivo e o total de sentimento (positivo e negativo). Com isto, podem ser medidas a polaridade e subjetividade relativa a uma dada entidade, ou em relação ao conjunto de todas entidades (denominado mundo). A intuição é que determinadas épocas (e.g. natal, eleições) implicam que mais sentimentos sejam expressos. A evolução destas métricas pode ser acompanhada visualmente em um gráfico.

Os autores ainda fazem uma comparação entre o conteúdo de blogs e jornais para as mesmas entidades, concluindo que o sentimento relacionado a certas entidades pode diferenciar-se em notícias e blogs, devido ao viés do meio de publicação.

O método foi validado utilizando indicadores externos existentes (e.g. sentimento detectado sobre o presidente, e pesquisas de intenção de votos) mostrando que os resultados são bastante próximos.

1.7.2.2. *Análise de sentimentos em notícias*

O foco do trabalho de Balahur *et al.* [Balahur et al. 2009b, Balahur et al. 2010] é a identificação de conteúdo subjetivo em notícias. Exceto por jornais sensacionalistas, este tipo

de texto evita expressar explicitamente opiniões, com objetivo de manter a seriedade e a suposta neutralidade da notícia. Opiniões neste meio são expressas de formas bem mais sutis, e que são linguisticamente difíceis de serem identificadas: argumentações tendenciosas; omissão ou destaque de fatos em detrimento de outros; citações de outras pessoas que expressam opiniões (e.g. "reiniciar este julgamento é um verdadeiro absurdo", disse o Ministro da Justiça); entre outros.

Dada a complexidade da análise de conteúdo implícito, a análise da subjetividade restringe-se a opiniões explicitamente expressas em citações, nas quais o conteúdo potencialmente de opinião é associado a um autor. Balahur *et al.* enfatizam a complexidade da mineração de opinião neste tipo de texto, e a restrição a citações como um primeiro passo para a compreensão dos diferentes problemas envolvidos. Este trabalho caracteriza-se por: a) nível de análise em citação, composta de uma ou mais sentenças; b) abordagem orientada a léxicos; c) uso de menções específicas para identificação de entidades.

A principal contribuição destes trabalhos é uma melhor caracterização da tarefa de mineração de opiniões em citações, através da realização de experimentos. As tarefas são definidas como:

- a definição do alvo da opinião: mesmo restringindo a citações, o alvo da opinião nem sempre é claro e explícito, podendo existir mais de um alvo. Utilizam-se entidades específicas como alvo, que estão em uma dada janela de distância das citações. As entidades são termos de busca de notícias (e.g. tsunami, Obama);

- a separação entre conteúdo ruim/bom, da opinião positiva ou negativa sobre este: a abordagem é baseada em léxicos, dos quais são removidos termos específicos que podem representar fatos com conotação negativa/positiva (e.g. crise, desastre, carnaval). Nos experimentos, foram utilizados rótulos (*tags*) que caracterizam categorias de notícias, e que possuem um sentimento associado. Os autores apontam que outras abordagens poderiam ter sido utilizadas com melhores resultados, inclusive escolha manual de termos.

- interpretação da polaridade sem informação contextual: os autores discorrem que em notícias existe uma grande distinção entre a opinião do autor, expressa na citação, e a opinião do leitor, de acordo com sua interpretação dos fatos. Esta divergência é uma grande dificuldade para o consenso em processos de anotação. Eles focam na classificação do primeiro tipo, usando apenas as informações explicitamente presentes no texto. Para a polarização, que foi baseada na abordagem de dicionário, foram experimentados 4 léxicos distintos, utilizados para polarizar todos os termos de sentimento encontrados na citação a uma dada distância do alvo. Diferentes abrangências de janela foram testadas.

Os experimentos envolveram 1.292 citações sobre as quais houve concordância entre os anotadores sobre a polaridade e o alvo. O melhor resultado (86% de acurácia) foi obtido com o uso combinado de dois léxicos (entre eles JRCTonality, desenvolvido pelo próprio grupo) e uma janela de 6 palavras entre alvo e termo de sentimento. Na análise de erros, os principais problemas detectados foram as citações polarizadas erroneamente como

neutras (devido à falta de palavras explícitas), o uso de ironias, emprego de abstrações para referir-se ao alvo, e múltiplos alvos.

1.7.3. Mídia Social

A mineração em mídias sociais têm a informalidade deste tipo de meio como um dos seus principais desafios. Não apenas o vocabulário pode ser bem específico e volátil, como o número de erros de digitação, de ortografia ou gramaticais pode invalidar a contribuição de análises linguísticas. Por outro lado, o volume da dados gerados sobre cada tópico é tão grande, quando comparado com outras fontes, que tais erros podem não ser relevantes. Um dos focos de trabalhos de mineração de opinião no Twitter, é o da capacidade de previsão, justamente com base neste grande volume. Dois destes trabalhos [Asur and Huberman 2010] [Bollen et al. 2011] são melhor detalhados no restante desta seção.

1.7.3.1. Previsão de indicadores de rentabilidade de filmes

A popularidade do Twitter é tal que organizações têm utilizado esta plataforma para divulgação e marketing de suas marcas e produtos. Este é o caso de filmes, onde produtores têm investido massivamente em publicidade e marketing voltado aos usuários do Twitter. Asur *et al.* em [Asur and Huberman 2010] realizam experimentos para verificar se o interesse que um filme desperta, particularmente na época de pré-lançamento, correlaciona-se com indicadores econômicos deste domínio. Mais especificamente, a partir do volume de *tweets* e do sentimento neles expressos, deseja-se prever: a arrecadação da primeira semana de bilheteria; os preços dos índices do *Hollywod Stock Exchange* (HSX); e o rendimento de todos os filmes de uma semana específica. Este trabalho caracteriza-se por: a) análise em nível de documento (*tweet*); b) uso de termos pré-definidos para determinação de entidades alvo; c) uso da abordagem de aprendizagem de máquina para polarização, e d) definição de métricas de agregação de sentimento utilizadas para previsão.

A proposta consiste de um estudo de caso onde foram analisados *tweets* de filmes específicos, sobre um período de três meses. Palavras-chaves, como URL's de material promocional e o título dos filmes, foram utilizadas para extrair os *tweets* relevantes e identificar os alvos. Juntamente com o conteúdo de cada *tweet*, também extraiu-se a data e hora de postagem, e o respectivo autor.

Os autores comparam dois tipos de previsão: quantitativo (quantidade de *tweets* e *retweets* contendo URL's de material promocional sobre o filme), e baseado em opinião expressa nos *tweets*. A segunda implica a necessidade de classificação da polaridade dos *tweets*, que foi realizada utilizando o classificador DynamicLMClassifier, disponível no pacote análise linguístico LingPipe[19]. Este classificador é ternário, i.e. classifica os *tweets* como positivo, negativo ou neutro. Os dados de treino foram rotulados manualmente utilizando-se o Amazon Mechanical Turk. Somente *tweets* com polaridade classificada de forma unânime pelos anotadores foram utilizados para treino. A acurácia do classificador foi elevada (98%) com *features* representando 8-gramas.

Para a previsão baseada em número de *tweets*, os autores definem a métrica taxa de

[6] www.alias-i.com/lingpipe

tweets de cada filme por hora (Equação 4). Esse métrica permite criar uma série temporal de menções de cada filme, e correlacioná-la usando regressão linear com as variáveis alvo, ou seja, previsão de bilheteria e previsão de valor dos índices HSX. Uma forte correlação foi encontrada (R^2 ajustado de 0,8, e de 0,9, respectivamente), significando alto poder de previsão.

$$\textit{Taxa-de-tweets(filme)} = \frac{|\textit{tweets(filme)}|}{|\textit{tempo (em horas)}|} \tag{4}$$

Já para a previsão baseada em sentimento, foram definidas duas outras métricas: a razão entre o total de *tweets* positivos e negativos e o total de *tweets* neutros, chamada pelos autores de *subjetividade*; e a razão entre *tweets* positivos e negativos. Os resultados obtidos foram inferiores aos obtidos com a métrica quantitativa de taxa de *tweets* por hora. No entanto, quando as métricas de sentimento são associadas à da taxa de *tweets*, houve uma melhoria no poder de predição.

É interessante notar que resultados semelhantes foram encontrados no domínio político, onde experimentos revelaram que a quantidade de *tweets* tiveram maior poder preditivo sobre o resultado de eleições, que o sentimento ou emoção neles expressos [O'Connor et al. 2010, Tumasjan et al. 2010].

1.7.3.2. Previsão do comportamento da bolsa de valores

O objetivo deste trabalho é semelhante ao da seção anterior, no domínio da bolsa de valores. Bollen *et al.* [Bollen et al. 2011] realizam experimentos para verificar se sentimento expresso no Twitter, chamado no trabalho de humor, tem influência sobre a bolsa de valores, e pode ser utilizado para prever seu comportamento usando o índice Dow Jones. Este trabalho caracteriza-se por: a) análise em nível de documento (*tweet*); b) uso de expressões específicas para determinar conteúdo de sentimento; c) uso da abordagem de léxica para polarização, onde um léxico de emoções também foi usado, e d) definição de métricas de agregação de sentimento utilizadas para previsão.

Foi realizado um estudo de caso envolvendo *tweets* correspondentes a um período de onze meses (Fevereiro, 2008 - Dezembro, 2008). A fase de identificação selecionou apenas *tweets* que continham sentimento explícito. Como critério, foram utilizadas expressões pré-definidas como "eu sinto", "eu estou sentindo", "eu não sinto", etc.

A classificação envolveu a abordagem léxica, mas dois tipos de classificação foram feitos: a) polaridade de sentimento (positivo e negativo), utilizando o léxico OpinionFinder [Wiebe and Riloff 2005]; e b) classificação da emoção (vide Seção 1.2.1), utilizada uma ferramenta denominada Google-Profile of Mood States (GPOMS). GPOMS analisa e classifica a emoção em seis diferentes dimensões: *calma, alerta, certeza, vitalidade, gentileza* e *felicidade*.

A polaridade da opinião foi sumarizada através de uma métrica que representa a razão entre a quantidade de termos positivos encontrados nos *tweets* e a de termos negativos, em um determinado dia. Já cada dimensão da emoção GPMOS foi totalizada por dia. Para todas estas métricas, foram criadas séries temporais, as quais foram comparadas com

eventos conhecidos do mesmo período: eleições e Ação de Graças. Exceto pela *gentileza* e *alerta* em relação às eleições, foi demonstrado que as técnicas utilizadas caracterizavam o humor típico destas datas. Uma análise estatística mostrou ainda correlação entre a polaridade da opinião e as dimensões de emoção *certeza*, *vitalidade*, e *alegria*, mas não com as demais.

Finalmente, foi usado um método baseado em redes neurais fuzzy (*self-organizing fuzzy neural network* - SOFNN) para correlacionar as séries temporais das polaridades das opiniões e das emoções, com a série temporal Down Jones Industrial Average (DIJA), e desenvolver um modelo preditivo. A abordagem utilizando o OpinionFinder não obteve bons resultados. No entanto, as emoções *calma* e *felicidade* demonstraram correlação com o DIJA em certos trechos do período analisado. O sentimento de calma foi o que obteve um melhor poder preditivo com um atraso (*lag*) de 3 dias. No entanto, o sentimento de calma só conseguiu prever corretamente o DIJA em períodos em que não há eventos inesperados (e.g. anúncio da Reserva Federal Americana). Isto significa dizer que consegue prever quando os índices se mantêm devido à falta de eventos externos importantes. Os autores concluem que a polaridade da opinião pode ser muito abrangente, e esconder aspectos cobertos pela subjetividade das emoções.

1.8. Conclusões e Direções Futuras

A mineração de opiniões é uma área de crescente interesse. Neste capítulo, discutimos seus conceitos básicos, os desafios na detecção de sentimento e de seu alvo, e técnicas que podem ser usadas para identificar, classificar a polaridade e agregar o sentimento expresso. Um estudo de caso foi usado para mostrar, em um caso real, como aplicar as técnicas discutidas, e os desafios e decisões envolvidos. Alguns trabalhos clássicos envolvendo diferentes fontes de opiniões foram apresentados a fim de discutir suas diferenças. O volume crescente de conteúdo subjetivo disponível diariamente, em particular nas redes sociais, motiva o crescimento da área com novas técnicas capazes de processar automaticamente textos, de forma escalável, robusta, precisa e independente de domínio e de linguagem. Muitas são as aplicações centradas na sumarização e visualização do sentimento, ou na predição de comportamentos com base no sentimento existente. Empresas, eventos (e.g. Olimpíadas), personalidades, estão interessadas na compreensão de como são percebidas pelo público em geral em tempo real, e nas mais variadas mídias.

A área ainda apresenta muitos problemas e oportunidades. Muitos esforços estão voltados à detecção de outros tipos de opinião: comparativa, implícita, dependente do observador, contradições, spams, ironias e sarcasmos, etc. A identificação de alvos e opiniões em mídias sem grau de estruturação, como notícias, é também outra área bastante importante. O desenvolvimento de recursos multilíngues permitirão o avanço do estado da arte, permitindo tratar corpus para os quais hoje não existem recursos (e.g. dados rotulados, léxicos, recursos para tratamento da língua natural, etc.). A evolução das técnicas de classificação para métodos escaláveis, menos sensíveis ao contexto ou a ruídos, e que combinam abordagens já existentes em um *framework* único é uma outra meta. Novas aplicações devem estabelecer soluções para *streaming* de dados, apoio a decisão baseado em sentimento, predições, entre tantas outras.

Referências

[Aggarwal and Zhai 2012] Aggarwal, C. C. and Zhai, C. (2012). *Mining text data*. Springer.

[Archak et al. 2007] Archak, N., Ghose, A., and Ipeirotis, P. G. (2007). Show me the money!: deriving the pricing power of product features by mining consumer reviews. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 56–65. ACM.

[Asur and Huberman 2010] Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 1, pages 492–499. IEEE.

[Baccianella et al. 2010] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, volume 10, pages 2200–2204. ELRA.

[Balahur et al. 2009a] Balahur, A., Kozareva, Z., and Montoyo, A. (2009a). Determining the polarity and source of opinions expressed in political debates. In *Computational Linguistics and Intelligent Text Processing*, pages 468–480. Springer.

[Balahur et al. 2009b] Balahur, A., Steinberger, R., Goot, E. v. d., Pouliquen, B., and Kabadjov, M. (2009b). Opinion mining on newspaper quotations. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and International Conference on Intelligent Agent Technology (WI-IAT)*, volume 3, pages 523–526. IET.

[Balahur et al. 2010] Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., and Belyaeva, J. (2010). Sentiment analysis in the news. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, volume 10, page 2216.

[Bollen et al. 2011] Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

[Bruce and Wiebe 1999] Bruce, R. F. and Wiebe, J. M. (1999). Recognizing subjectivity: a case study in manual tagging. *Natural Language Engineering*, 5(2):187–205.

[Calais Guerra et al. 2011] Calais Guerra, P. H., Veloso, A., Meira Jr, W., and Almeida, V. (2011). From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 150–158. ACM.

[Carvalho et al. 2009] Carvalho, P., Sarmento, L., Silva, M. J., and de Oliveira, E. (2009). Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, pages 53–56. ACM.

[Castellanos et al. 2011] Castellanos, M., Dayal, U., Hsu, M., Ghosh, R., Dekhil, M., Lu, Y., Zhang, L., and Schreiman, M. (2011). LCI: a social channel analysis platform for live customer intelligence. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 1049–1058. ACM.

[Chen and Wang 2014] Chen, L. and Wang, F. (2014). Sentiment-enhanced explanation of product recommendations. In *Proceedings of the 23rd International World Wide Web Conference(WWW) - Companion Volume*, pages 239–240. ACM.

[Dave et al. 2003] Dave, K., Lawrence, S., and Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, pages 519–528. ACM.

[de Paiva et al. 2012] de Paiva, V., Rademaker, A., and de Melo, G. (2012). OpenWordNet-PT: An open brazilian wordnet for reasoning. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 353–360.

[Dey and Haque 2009] Dey, L. and Haque, S. (2009). Opinion mining from noisy text data. *International journal on document analysis and recognition*, 12(3):205–226.

[Fan and Chang 2010] Fan, T.-K. and Chang, C.-H. (2010). Sentiment-oriented contextual advertising. *Knowl. Inf. Syst.*, 23(3):321–344.

[Fayyad et al. 1996] Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Commun. of the ACM*, 39(11):27–34.

[Fellbaum 2010] Fellbaum, C. (2010). *WordNet*. Springer.

[Ghani et al. 2006] Ghani, R., Probst, K., Liu, Y., Krema, M., and Fano, A. (2006). Text mining for product attribute extraction. *ACM SIGKDD Explorations Newsletter*, 8(1):41–48.

[Godbole et al. 2007] Godbole, N., Srinivasaiah, M., and Skiena, S. (2007). Large-scale sentiment analysis for news and blogs. In *Proceedings of the First International AAAI Conference on Weblogs and Social Media (ICWSM)*, volume 2.

[Hu and Liu 2004] Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 168–177. ACM.

[Indurkhya and Damerau 2010] Indurkhya, N. and Damerau, F. J. (2010). *Handbook of Natural Language Processing, 2nd Ed.* CRC Press.

[Jurafsky and Martin 2010] Jurafsky, D. and Martin, J. H. (2010). *Speech and Language Processing, 2nd Ed.* Prentice Hall.

[Ku et al. 2006] Ku, L., Liang, Y., and Chen, H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, volume 100107. AAAI Press.

[Liu 2010] Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, 2:627–666.

[Liu 2012] Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

[Liu et al. 2013] Liu, Q., Gao, Z., Liu, B., and Zhang, Y. (2013). A logic programming approach to aspect extraction in opinion mining. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and International Conference on Intelligent Agent Technology (WI-IAT)*, pages 276–283.

[Mohammad et al. 2013] Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the Seventh International Workshop on Semantic Evaluation Exercises (SemEval)*.

[Ng 2010] Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1396–1411. The Association for Computer Linguistics.

[O'Connor et al. 2010] O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM)*, pages 122–129. AAAI Press.

[Pak and Paroubek 2010] Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, volume 10, pages 1320–1326. ELRA.

[Pang and Lee 2008] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

[Pang et al. 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

[Potts 2011] Potts, C. (2011). Sentiment analysis tutorial. In *Sentiment Analysis Symposium 2011 - Tutorial Notes*. Online. Captured at: sentiment.christopherpotts.net/overview.html, May 2014.

[Qiu et al. 2011] Qiu, G., Liu, B., Bu, J., and Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1):9–27.

[Sarawagi 2008] Sarawagi, S. (2008). Information extraction. *Foundations and trends in databases*, 1(3):261–377.

[Sarmento et al. 2009] Sarmento, L., Carvalho, P., Silva, M., and de Oliveira, E. (2009). Automatic creation of a reference corpus for political opinion mining in user-generated content. In *Proceedings of the 1st international CIKM Workshop on Topic-sentiment analysis for mass opinion*, pages 29–36. ACM.

[Silva et al. 2012] Silva, M., Carvalho, P., and Sarmento, L. (2012). Building a sentiment lexicon for social judgement mining. *Computational Processing of the Portuguese Language*, pages 218–228.

[Sohail et al. 2013] Sohail, S. S., Siddiqui, J., and Ali, R. (2013). Book recommendation system using opinion mining technique. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1609–1614. IEEE Press.

[Souza et al. 2011] Souza, M., Vieira, R., Busetti, D., Chishman, R., and Alves, I. M. (2011). Construction of a portuguese opinion lexicon from multiple resources. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL)*. SBC.

[Stone et al. 1966] Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). *The General Inquirer: A Computer Approach to Content Analysis.* MIT Press.

[Strapparava and Valitutti 2004] Strapparava, C. and Valitutti, A. (2004). Wordnet affect: an affective extension of wordnet. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, volume 4, pages 1083–1086. ELRA.

[Tan et al. 2006] Tan, P.-N., Steinbach, M., and Kumar, V. (2006). *Introduction to data mining*. Addison Wesley.

[Tang et al. 2009] Tang, H., Tan, S., and Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Syst. Appl.*, 36(7):10760–10773.

[Tausczik and Pennebaker 2007] Tausczik, Y. R. and Pennebaker, J. W. (2007). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

[Thet et al. 2010] Thet, T., Na, J., and Khoo, C. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 36(6):823–848.

[Tsytsarau and Palpanas 2012] Tsytsarau, M. and Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478–514.

[Tumasjan et al. 2010] Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM)*, pages 178–185. AAAI Press.

[Tumitan and Becker 2013] Tumitan, D. and Becker, K. (2013). Tracking sentiment evolution on user-generated content: A case study in the brazilian political scene. In *Proceedings of the 28th Brazilian Symposium on Databases (SBBD)*, pages 135–144. SBC.

[Tumitan and Becker 2014] Tumitan, D. and Becker, K. (2014). Sentiment-based features for predicting election polls: a case study on the brazilian scenario. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and International Conference on Intelligent Agent Technology (WI-IAT)*, pages 1–8. IEEE Computer Society.

[Turney 2002] Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424. Association for Computational Linguistics.

[Wiebe and Riloff 2005] Wiebe, J. and Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Computational Linguistics and Intelligent Text Processing*, pages 486–497. Springer.

[Wiebe et al. 2005] Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.